SOMDE: A scalable method for identifying spatially variable genes with self-organizing map

Minsheng Hao[1], Kui Hua[1*] and Xuegong Zhang[1,2*]

[1] MOE Key Laboratory of Bioinformatics and Bioinformatics Division, BNRIST, Department of Automation, Tsinghua University, Beijing 100084, China

[2] School of Life Sciences, Center for Synthetic and Systems Biology, Tsinghua University, Beijing 100084, China

[*]Correspondence：zhangxg@tsinghua.edu.cn, stevenhuakui@gmail.com

Abstract

Recent developments of spatial transcriptomic sequencing technologies provide powerful tools for understanding cells in the physical context of tissue micro-environments. A fundamental task in spatial gene expression analysis is to identify genes with spatially variable expression patterns, or spatially variable genes (SVgenes). Several computational methods have been developed for this task. Their high computational complexity limited their scalability to the latest and future large-scale spatial expression data.

We present SOMDE, an efficient method for identifying SVgenes in large-scale spatial expression data. SOMDE uses self-organizing map (SOM) to cluster neighboring cells into nodes, and then uses a Gaussian Process to fit the node-level spatial gene expression to identify SVgenes. Experiments show that SOMDE is about 5-50 times faster than existing methods with comparable results. The adjustable resolution of SOMDE makes it the only method that can give results in ~5 minutes in large datasets of more than 20,000 sequencing sites. SOMDE is available as a python package on PyPI at https://pypi.org/project/somde.

1. Introduction

The spatial location of cells in tissues and corresponding gene expression profiles play pivotal roles in the study of tissue mechanism and tumor immune microenvironment. Spatial transcriptomics sequencing technologies provide gene expression profiles with spatial information, filling the gap between high throughput transcriptomics and their spatial context.

Since the publication of Spatial Transcriptomic (ST) (Ståhl *et al.*, 2016) to today's spatial transcriptomic sequencing technologies, the number of measured spatial data sites in one sample has increased from a few hundred to tens of thousands. Targeted in-situ sequencing technologies such as seqFISH+ (Eng *et al.*, 2019) and MERFISH (Moffitt *et al.*, 2018) capture single-cell level transcripts under the camera field of view (FOV). They splice multiple adjacent FOV results to form datasets with a large number of cells. Advanced Spatial Transcriptomic methods Slide-seq (Rodriques *et al.*, 2019) and 10X Visium (the commercial version of original ST) can sequence 5,000~25,000 spatial data sites with a resolution of 10 or 55 μm, respectively. It is foreseeable that the scale of spatial transcriptomic data will increase quickly.

Spatial transcriptomic data provides gene expressions with physical location information. Spatial variations of gene expression reflect the cell-cell interaction relationship (Dries *et al.*, 2019) and help determine compositions of cell types that perform spatial specific functions. Identifying spatially variable genes (SVgenes) is the basic task for spatial transcriptomic data analysis. Compared with previous tasks of identifying highly variable genes (HVG) from gene expression profiles, SVgene identification needs to consider not only variations between cells but also the spatial significance of gene expression.

Several methods have been proposed for identifying SVgenes based on early spatial transcriptomic data. For example, Trendsceek (Edsgärd *et al.*, 2018), SpatialDE (Svensson *et al.*, 2018) and SPARK (Sun *et al.*, 2020) find SVgenes via the statistical spatial correlation models. scGCO (Zhang *et al.*, 2018) detects SVgenes with segmented spatial expression patterns. SilhouetteRank (Dries *et al.*, 2019) identifies SVgenes whose binarized gene expression has clustered patterns in space. Despite the great success of those methods in low-throughput spatial transcriptomic data, the high computational complexity hinders their application in large-scale datasets. For example, when the number of sites exceeds 20,000, the first three methods will require at least 1000 minutes to get results (Sun *et al.*, 2020).

To better cope with the growth of spatial transcriptomic data sizes, we present SOMDE, a scalable method to identify SVgenes with high computational efficiency. The method uses the self-organizing map (SOM) neural network and Gaussian process to model spatial data. We conducted a series of experiments in five representative datasets obtained by two scalable protocols. Results showed that SOMDE gave similar SVgenes as existing methods, but in much less time with enhanced spatial pattern visualization. The adjustable number of nodes enables SOMDE to get results in large-scale datasets with more than 20,000 sites in only 5 minutes. We have developed SOMDE as a python package available on PyPI at https://pypi.org/project/somde.

2. Methods

2.1 SVgene identification

We call a gene an SVgene if its transcript expression is highly correlated with the spatial location. The spatial expression pattern of SVgenes depends on the spatial histological structure or other spatial factors, and exhibits clustered, periodic or other characteristics in space.

SVgene identification is a fundamental task in spatial gene expression data analysis. Existing methods models spatial expression variability from multiple perspectives. Trendsceek, SpatialDE and SPARK (Edsgärd *et al.*, 2018, Svensson *et al.*, 2018, Sun *et al.*, 2020) consider the correlation between the distribution of gene spatial expression and the data sites locations via statistical models. scGCO (Zhang *et al.*, 2018) applies Delaunay triangulation across the tissue to generate a sparse graph representation of data sites. SilhouetteRank (Dries *et al.*, 2019) proposed in Giotto Analyzer binarizes gene expression and divides spatial sequencing sites into two clusters.

The above existing methods have succeeded in the early spatial data. Trendsceek (Edsgärd *et al.*, 2018) assesses the significant dependency between gene expression levels and spatial locations. SpatialDE (Svensson *et al.*, 2018) decomposes gene expression variability into spatial and non-spatial parts and identifies SVgenes using statistical testing. In 2020, Sun et al. presented spatial pattern recognition via kernels (SPARK) (Sun *et al.*, 2020). It detects SVgenes through a calibrated p-value calculation based on a generalized linear spatial model. scGCO (Zhang *et al.*, 2018) adopts binary cuts on the graph via the optimization of Markov Random Fields. SilhouetteRank (Dries *et al.*, 2019) introduces the silhouette score as a metric and identifies SVgenes with spatial aggregation expression patterns.

All these methods have been challenged by the latest large-scale datasets. The computational complexity of identifying SVgenes increases significantly with the growth in the number of spatial sequencing sites. For the first three methods, the size of covariance kernel matrices in their model grows quadratically as the number of spatial sites increases, and therefore are both time and memory consuming to be optimized. The latter two methods are relative faster due to the introduce of the binarization step and the application of

non-statistical modeling. However, the identification result is sensitive to the binarization threshold and an appropriate threshold needs multiple attempts. The exploration and verification of every selected threshold are still time-consuming.

2.2 Our approach

The key idea of SOMDE is to construct a condensed representation of spatial transcriptomic data that both preserves the information of SVgenes and reduce the downstream computational complexity. We use the Self-Organizing Map (SOM) neural network model to adaptively integrate neighboring data into different nodes, and then identify SVgenes based on the node-level spatial location and gene expression information using a modified Gaussian process (Fig.1).

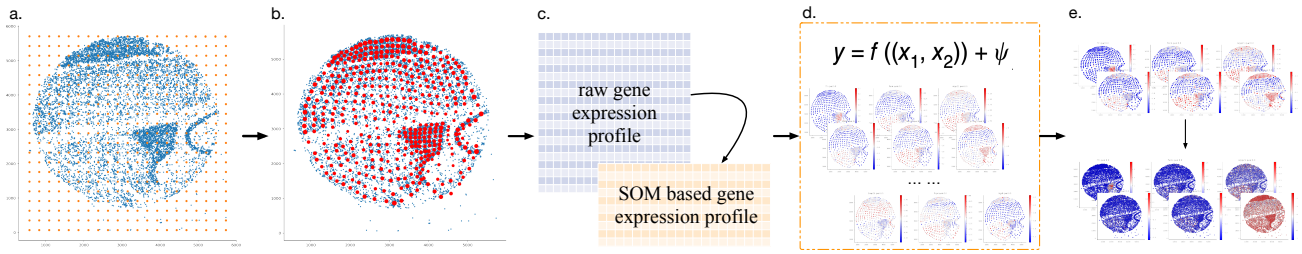

Fig. 1. A schematic overview of SOMDE. (a) SOMDE initializes a self-organizing map (SOM) on the tissue spatial domain. (b) By training the SOM with data site locations, SOMDE merges the original data sites (blue spots) into different nodes (red spots). (c) Then SOMDE converts the original spatial gene expression to node-level gene meta-expression profiles. (d) SOMDE models the condensed representation of the original spatial transcriptome data with a modified Gaussian process to quantify the relative spatial variability. (e) SOMDE identifies genes with high spatial variability expression patterns as SVgenes, and maps node-level expression patterns of SVgenes to their original expression patterns.

2.2.1 Data site integration

The spatial expression of SVgenes usually has certain continuity. A proper strategy of data site integration should preserve the spatial expression patterns of SVgenes and the topological structure of data sites. SOM meets both the requirements. SOMDE first adopts SOM to integrate spatial data sites into nodes and then assigns the gene expression profile of each node.

SOM is an unsupervised neural network first proposed by Kohonen (Kohonen, 1982). It is an array of neurons with weights connected to all input dimensions and all neurons are arranged in a spatial grid. There is no direct connection between SOM neurons, but during training adjacent neurons on the grid will affect each other. After proper training, SOM can form an organized map in the 2D node space for the input data. The map preserves the topological constraints and density relations among the input data in the original space (Uriarte et al. 2005). When the input data dimension is also 2D, the SOM forms a condensed map with a small number of nodes according to the topology and density among the input data. It can be taken as an adaptive down-sampling of the input space that well preserves the original spatial information

We build the SOM by taking spatial locations $\boldsymbol{x} = (x_1, x_2, \dots, x_C)$ of the sequencing data sites as the input. C is the total number of data sites. The map size of a $N \times N$ SOM is chosen according to the formula:

$$N = \sqrt{\frac{C}{k}} \tag{1}$$

where k is the neighbor number. It donates the expected average number of original data sites each SOM node represents. The same hyperparameter k maintains the same sparseness of SOM nodes and the spatial sites, regardless of the size of the dataset.

We initialize the SOM node weights with the uniform grid coordinates in the space. SOM adjusts the weight vector of each node through competitive training. The traditional training methods demand to set the initial value and attenuation function of the learning rate, which causes the instability of the training results. We applied the batch SOM algorithm (Wittek *et al.*, 2017) that parallelly updates the node weights after each epoch to solve this issue. The updating formula for corresponding weight vector m_i of node i is:

$$m_i(t_f) = \frac{\sum_{t'=t_0}^{t'=t_f} h_{c,i}(t')x(t')}{\sum_{t'=t_0}^{t'=t_f} h_{c,i}(t')}, \tag{2}$$

where $h_{c,i}(t)$ is the neighborhood function between the node c and node i. The t_0 and t_f represent the beginning and end of the training epoch, respectively. This batch training form avoids the needs for setting the hyperparameters, and all weights are updated at the same time after one epoch. Each update of one SOM node only affects its neighboring nodes, so the training process has lower computational complexity. It takes less than one second to train SOM in a dataset of 20,000 sites. Since the input data dimension is the same as the node, the value of trained node weight vectors m_i is directly considered as the spatial coordinates $\tilde{x}_i$ of node i. The spatial locations $\tilde{\boldsymbol{x}} = (\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{N\times N})$ of all nodes compose the sparse topology of the original data.

After the SOM training, each data site maps to a unique node in the SOM plain. Each SOM node represents the group of sites that map to it. To represent the expression of a gene in this group, we define the "meta-expression" of the gene at a SOM node as the linear combination of the max value and average value of the gene expression in the group of sites that the node represents. For example, $x_{S1}, x_{S2}, \dots$ are a set of neighboring data sites mapping to a SOM node (i,j), the meta-expression $\tilde{y}_{i,j}$ of one gene at this node is

$$\tilde{y}_{i,j} = \gamma \cdot max\,(y_{S1}, y_{S2}, \dots) + (1-\gamma) \cdot \mathrm{a}vg(y_{S1}, y_{S2}, \dots) \tag{3}$$

where y_{Si} is the expression value of the gene at data site x_{Si}. The combination ratio γ should be adjusted according to the sparseness and quality of the data. We use $\gamma = 0.5$ in current experiment.

With this mapping and data integration for all gene and all data sites, we obtain the condensed representation of the original spatial transcriptome data by the meta-expression map in the SOM nodes plain that best preserves the original topological and expression information. The resolution of this condensed spatial transcriptomic map can be adjusted by the hyperparameter k.

2.2.2 Spatial expression variability identification

The secondary step is identifying gene spatially expression variability with the condensed spatial transcriptomic map. We consider one gene meta-expression at one time and introduces an adjusted Gaussian process H_G in SOMDE to model the spatial correlation. H_G decomposes the expression variability into spatial and non-spatial components:

$$p(\tilde{\boldsymbol{y}} \mid H_G, \tilde{\boldsymbol{x}}, \Theta) = \mathcal{N}\left(\tilde{\boldsymbol{y}} \,\middle|\, \mu \cdot \mathbf{1},\ \sigma_s^2 \cdot \boldsymbol{\Sigma}_{k(\tilde{x},\tilde{x}'|\theta)} + \delta \cdot \boldsymbol{I}\right) \tag{4}$$

where $\tilde{\boldsymbol{y}}$ donates one gene meta-expression in the SOM plain and $\tilde{\boldsymbol{x}}$ donates all SOM nodes locations. $\delta \cdot \boldsymbol{I}$ indicates the non-spatial variance given by Gaussian distributed noise in all observed gene meta-expression. $\boldsymbol{\Sigma}_{k(\tilde{x},\tilde{x}'|\theta)}$ captures the spatial variation and is the covariance function generated from a selected kernel function $k(\cdot)$. Here $\tilde{x}$ and $\tilde{x}'$ represent the position vectors of any two SOM nodes. The kernel function $k(\cdot)$ introduces the spatial coefficient and is usually chosen from the Gaussian, linear or periodic kernel. Θ contains all the parameters that need to be optimized including the mean value μ, variance σ_s^2 and noise variance δ.

The two separate parts in the model variance allow us to get the fraction of spatial variation (FSV) in total variance from calculating the proportion of σ_s^2 in $\sigma_s^2 + \delta$. Compared with the original Gaussian process model, H_G model assumes one gene meta-expression at all spatial locations has the same mean value μ so that the non-spatial variance cannot be regressed out. Thus the maximum likelihood value of H_G mainly depends on spatial variance. We directly take the likelihood value as the spatial variability score for each gene.

Since SOMDE takes SOM nodes as the basic units, the covariance matrix is reduced to $1/k^2$ of the original matrix compared with previous statistical methods. We applied gradient optimization to get the best estimation of parameters in H_G . Finally, SOMDE ranks all genes according to their spatial variability scores (from high to low) as the SVgene rank.

We use similar log ratio test as SpatialDE to determine the statistical significance of each gene's spatial expression variability. Consider a Gaussian model H_0 without the spatial covariance term：

$$p(\tilde{\boldsymbol{y}} \mid H_0, \tilde{\boldsymbol{x}}, \theta) = \mathcal{N}(\tilde{\boldsymbol{y}} \mid \mu \cdot \mathbf{1}, \delta \cdot \boldsymbol{I}) \tag{5}$$

where μ and δ represent the mean value and noise variance, respectively. The log-ratio test between H_0 and H_G reflects the significance of the spatial variance. And the P-value can be estimated by assuming the log-likelihood ratios (LLR) are $\chi 2$ distributed with one degree of freedom.

The top SVgenes ranks given by SOMDE at different resolutions may be different. To get a unified result, we propose a k-free solution called combined-SOMDE (cSOMDE) to merge all results. For each gene, cSOMDE takes the geometric mean of ranking results given by SOMDE with three different k as the final rank. This strategy considers the spatial variability of the same gene expression at different reduced resolutions.

For the implementation, SOMDE is implemented in Python and uses somoclu V1.7.5 for multi-thread building SOM. We choose SpatialDE V1.1.0 and Giotto V0.3.1 as representatives of statistical methods and non-statistical methods for comparison. All experiments are based on 16 AMD Ryzen 7 1700X Eight-Core Processors.

3. Results

3.1 Datasets and Experiments

We applied SOMDE to five large-scale datasets: 10X Brain, Hippocampus (Hipp), near Hippocampus (nHipp), Liver and Kidney. These datasets are produced by two scalable spatial sequencing protocols: The 10X Brain datasets is obtained by 10X Visium, and other four datasets are obtained by Slide-seq. Brief descriptions of these datasets are shown in Table 1.

The spatial resolution of the 10X Visium and Slide-seq protocol is $55\mu m$ and $10\mu m$, respectively. Although these data are not at the single-cell level, the measured gene expressions are still highly correlated with spatial locations. The large gene and data site number is suitable for verifying the computational efficiency of our SVgenes identification method.

We set the default neighbor number k to 20 for all five datasets without deliberate selection. SOMDE takes spatial information and normalized gene expression matrix as input, gives the SVgene rank and test the statistical significance of spatial variability with q-value. Genes with q-value smaller than 0.05 are identified as SVgenes. We further confirm the validity of identified SVgenes in nHipp and 10X Brain datasets from the biological and mathematical views. The SOMDE running time of all the above experiments has been recorded.

SOMDE and its variants were compared with the existing two methods in the 10X Brain and nHipp datasets. They are SpatialDE that uses the modified Gaussian process directly for identification, and Giotto (SilhouetteRank) that introduces the silhouette rank to sort

binarized expressed genes. For the cSOMDE method, we set neighbor number k=5, 20, 40 and k=4, 20, 40 in 10X Brain and nHipp datasets, respectively. Other methods use the default settings. Since not all methods can get the SVgenes number and these methods use different criteria for identifying gene spatial expression variability, we only compare the gene rank similarity.

Protocol	Tissue	Gene number	Site number
10X Visium	Brain	14,414	2,698
Slide-seq	Hippocampus	3,235	12,218
Slide-seq	near Hippocampus	2,555	9,650
Slide-seq	Liver	2,238	17,454
Slide-seq	Kidney	4,650	22,860

Table 1. The overview of datasets used in our experiments. Five datasets sequenced by two different protocols are used in our experiments. 10X Brain data is downloaded directly from official 10X Genomics websites. Slide-seq datasets are downloaded from SpatialDB (Fan *et al.*, 2019).

3.2 SOMDE results

SOMDE successfully gave the SVgene rank and the SVgenes number in each dataset (Fig. 2. & Supplementary Materials). In the 10X Brain dataset, SOMDE found 5,455 SVgenes. In the Liver dataset, the SVgenes number is the smallest with only 164 genes. SOMDE found 699, 379, 522 SVgenes in Hipp, nHipp and Kidney datasets, respectively. The top 5 SVgenes found by SOMDE in the 10X Brain dataset are *Sparc, Agt, Tcf7l2, Slc6a11* and *Nrgn*. In the nHipp and Hipp dataset, *Nrgn* and *Tcf7l2* were also identified as the top 10 SVgenes. *Ttr* gene has the most gene expression variability in both nHipp and Hipp datasets. In Liver and Kidney datasets, the top 5 SVgenes are *Mup17, Hpx, Mgst1, Glul, Mup3* and *Napsa, Kap, Aadat, Mpv17l, Acadm*. All SVgene rank lists of these 5 datasets are available in the supplementary material.

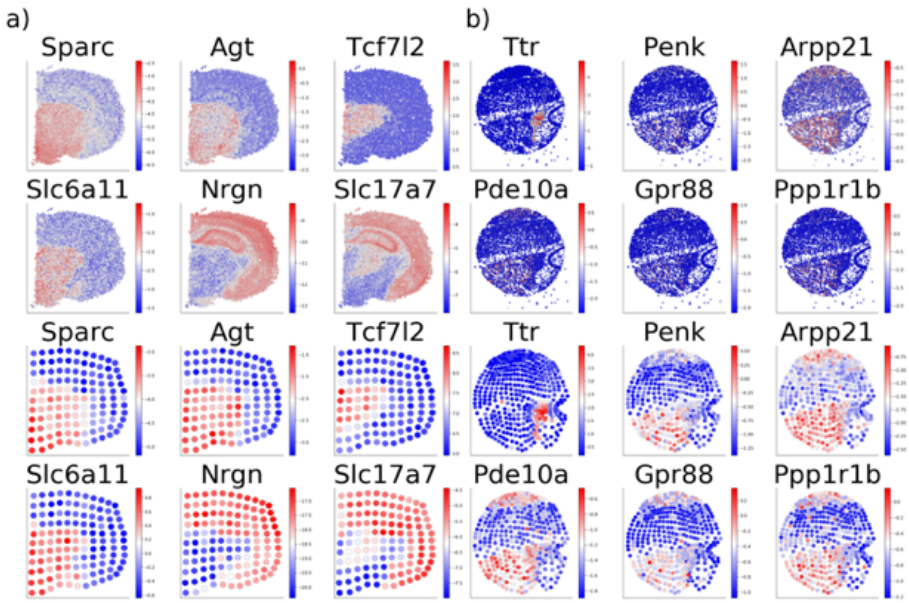

Fig. 2. The top 6 SVgenes found by SOMDE are visualized in original and SOM space. Top 6 SVgenes found in 10X Brain and nHipp datasets. Each subplot donates one gene spatial expression pattern. Color implies the relative expression value within the plot. The first two rows and the last two rows are in the original space and SOM space, respectively. (Zoom in for details)

From the results we found that the different proportions of SVgene among datasets are more likely due to the biological characteristics of the tissue, rather than technical bias or different data size. In tissues with complex spatial correlated functions such as the whole adult mouse brain or hippocampus, the proportion of SVgenes is far greater than in the adult mouse kidney or liver tissues.

We plotted the top 50 SVgenes spatial expression patterns at the original resolution and node-level resolution in all datasets. (Fig.2 & Supplementary Figure) Each spot in figure donates one data site or one SOM node, and the color represents the relative expression values within one plot. The visualization results on the original data and condensed map show the similar topological structure of data sites and spatial gene expression patterns.

The integration step of SOMDE improves the signal-to-noise ratio in the Slide-seq nHipp dataset. *Pde10a, Gpr88* and *Ppp1r1b* genes (Fig. 2b) have low expression levels in the original space so that it is hard to visualize their high spatial expression variability due to the noise and sparsity. After integration these gene expressions show strong spatial patterns and variabilities in our condensed map. SOMDE successfully denoises gene expressions via the combination of the maximum and average local expression.

We further analyze our SOMDE results in the Slide-seq nHipp and 10X Brain datasets to verify the identification results. We use the MGI database (Smith *et al.*, 2019) and Allen Brain Atlas (Lein *et al.*, 2007) as references to verify the genes' morphological function and spatial expression distribution. All SVgenes identified by SOMDE showed clear spatial variation and consistent with existing In Situ Hybridization (ISH) patterns of the same tissue. For example, *Nrgn*, one of the top five SVgenes in the 10X Brain dataset, is responsible for the transcription of neurogranin protein that encodes transcription and signal synapse transduction. Both 10X Brain and ISH data reveal that it mainly expressed in the cerebral cortex and hippocampal formation. *Camk2n1*, another identified SVgene identified in mouse brain, is responsible for the production of enzyme regulators and highly expressed on the outer and posterior parts of the cerebral cortex in our condensed map. This spatial pattern is also cross validated by the ISH data. Both *Nrgn* and *Camk2n1* play an important role in synaptic long-term potentiation (Ling, K.H., 2011), and they have the similar spatial variability score and SV gene rank, suggesting our results is of potential biological interest.

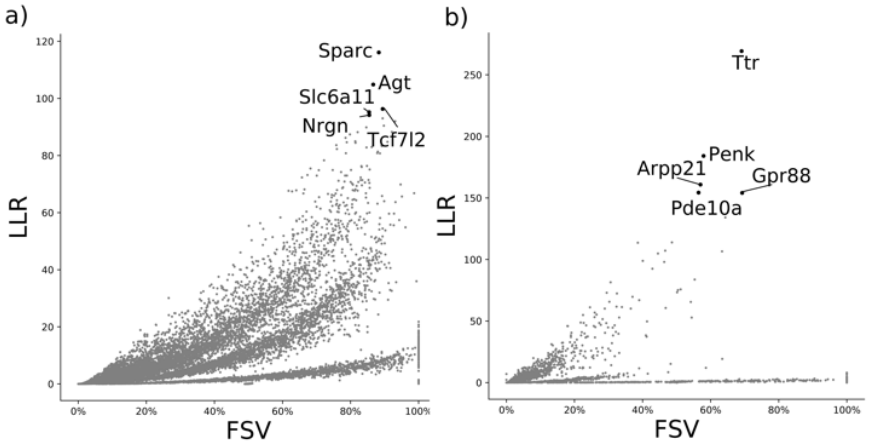

Fig. 3. Quantitative results of SOMDE in 10X Brain and nHipp datasets. The x-axis represents the fraction of variance explained by spatial variation (FSV), and the y-axis represents the log-likelihood ratio (LLR). The top 5 genes selected by our method are high on both indicators. Subfigure a) and b) show results in brain datasets and nHipp dataset.

We also confirm the validity of SOMDE from the mathematical view. Fig. 3 demonstrates the log-likelihood ratio (LLR) and the fraction of spatial variation (FSV) of all genes in 10X Brain and nHipp datasets. Each spot donates one gene and we highlight the spots

of the top 5 SVgenes. The top five SVgenes in both datasets identified by our method have high FSV and LLR values. Sparc gene has the maximum LLR value corresponding to the most significant spatial patterns shown in Fig.2. For all highly spatially variable genes, LLR and FSV values are positively correlated as expected. The distribution of all gene spots indicates most spatially variable genes have significant statistical values, which proves that the integration of spatial sites does not affect the identification of gene variability.

Table 2 shows the running time of SOMDE in all five datasets. Running time refers to the total time from training self-organizing map to obtaining all SVgene ranks and spatial variability scores. The adjustable map size guarantees the condensed maps are at the same resolution in different datasets. SOMDE gives all the results with no more than 5 minutes regardless of the gene and data site number. We can conclude that the condensed transcriptomic maps make the computational efficiency of SOMDE not limited by the dataset size.

Dataset	Gene number	Site number	Map	Time(s)
10X Brain	14,414	2,698	11×11	247
Hipp	3,235	12,218	24×24	56
nHipp	2,555	9,650	21×21	57
Liver	2,238	17,454	29×29	58
Kidney	4,650	22,860	33×33	147

Table 2. the running time of SOMDE in multiple datasets. The results of the SOMDE in different datasets with $k = 20$. Although the running time is affected by the gene and sequencing site number, SOMDE gives results in less than 5 minutes in multiple datasets (some datasets even exceed 20,000 data sites), which shows the superiority of our method in terms of running time.

3.3 Compared with existing methods

3.3.1 Gene ranking similarity

All four methods reveal similar judgments on the most spatially variable genes, regardless of their own criteria or principle. In the nHipp datasets, *Ttr, Penk, Arpp21* and *Enpp2* are accurately identified as the top10 SVgenes with the highest spatial variability by all methods. Among the top 100 SVgenes identified by SOMDE, cSOMDE, SpatialDE and Giotto identified the same 92, 67 and 51 SVgenes, respectively. Among the top 200 SVgenes of SOMDE, cSOMDE identified 186 identical SVgenes, SpatialDE identified 147 SVgenes, and 75 SVgenes are also in the Giotto top 200 results. Among the top 500 SVgenes identified by SOMDE, 457, 401 and 142 SVgenes are also in the top 500 places of cSOMDE, SpatialDE and Giotto respectively.

Similar results are obtained in the 10X Brain dataset. *Nrgn, Mef2c, Agt, Sparc, Camk2n1, Tcf7l2* and *Camk2a* are in the intersection of the top 20 SVgenes identified by all methods. 45, 41 and 87 SVgenes of the SOMDE identified top 100 SVgenes are also ranked in the top 100 by cSOMDE, SpatialDE and Giotto, respectively. Among the top 500 SOMDE SVgenes, cSOMDE identified 435 identical SVgenes, SpatialDE identified 274 identical SVgenes, and 277 SVgenes are also in the Giotto top 200 identification results. Compared with Giotto method, our method and SpatialDE use similar statistical models to infer the spatial variability. Thus SOMDE have a high consistency with the results of SpatialDE.

SOMDE and cSOMDE have the almost identical SVgene ranks in both two datasets, which indicates that the spatial expression variabilities of most genes are almost the same at adjacent different resolutions. It is in line with the similar expression patterns between original and node-level resolution shown in Fig.2. This result also proves that our method is not sensitive to the selection of neighbor number k.

3.3.2 Running time comparison

We selected three real representative datasets: 10X Brain, nHipp, and Kidney datasets to test the SOMDE performance in large scale datasets. The number of sequencing sites in these three datasets gradually increases, as shown in Table 2. The comparison results of four methods are shown in Fig. 4a. SOMDE took 247, 57 and 144 seconds to give results in each dataset. Giotto and SpatialDE took more than 5000 seconds in the Kidney dataset. The adjustable resolution of SOM makes SOMDE the only method that gets results in large-scale datasets in 300 seconds. Unlike other methods, the computational complexity of SOMDE does not increase with the growth of the dataset size, and is only affected by the number of genes. Therefore, the running time in Liver and Kidney datasets is even less than which in 10X Brain dataset.

Parameter $\boldsymbol{k}$ controls the size of the condensed map in SOMDE, thereby affecting the resolution of down sampling. A map with a coarser resolution can save time while losing more details. The previous results prove that SOMDE gives similar results in the adjacent resolution, and thus not a specific value but a proper interval of the parameter $\boldsymbol{k}$ is needed to balance the computation efficiency and identification effectiveness. In Fig. 4b, we gradually increased the number of neighbors k and recorded the running time in these three datasets. The running time has a sharp drop when the neighbor number first increases, but tends to stabilize as $\boldsymbol{k}$ further increases. We found that only a small reduction in the number of nodes brings the most significant computational improvement. The back part of the curve in Fig. 4b illustrates that our method successfully integrates space sites and improves computational efficiency with the condensed map, and the site number is no longer the main factor affecting the running time of the identification.

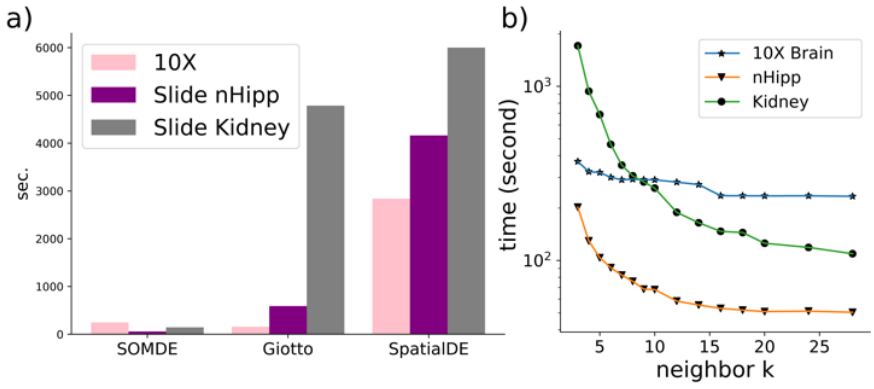

Fig. 4. Running time Comparison. a) The total running time of 3 methods in 3 datasets with different sizes. Colors correspond to different datasets. b) Running time of SOMDE under different neighbor number. SOMDE gives results rapidly in all three datasets and has the most impressive improvements in the kidney dataset which has the largest number of spatial sites.

We recommend $\boldsymbol{k}$=5-20 as the initial setting for SOMDE. From the experiments we found that the variability of spatial gene expression barely changes within 5-20 neighboring sites or cells, so the integration step with this setting does not weaken the identification of SVgene spatial expression patterns. The computational efficiency can theoretically be improved by 5 times compared with non-integration identification.

4. Discussion

Identifying spatially variable genes is a fundamental task for understanding spatial transcriptome data. The development of spatial transcriptomic sequencing technology brings an increase in the data size, which in turn challenges existing methods for their application to large-scale datasets. We present SOMDE, an SVgenes identification method that extends the application to the large-scale dataset by

combining the advantages of machine learning and statistical models. SOMDE integrates gene expression and spatial locations to a condensed map and identifies gene spatial expression variability via SOM and Gaussian Process.

Experiments in multiple datasets with different sequencing protocols prove that the SOM-based condensed map well preserved the topological structure and the gene spatial expression patterns. The identified SVgenes given by SOMDE are highly consistent with other methods and existing gene ISH data. Our method also enhances the spatial patterns of SVgene expressions, which provides great help for visualization.

The high SVgene rank similarity between SOMDE and cSOMDE results shows that adjacent different condensed map sizes do not significantly change the identification results. Therefore one SOMDE model is sufficient in most scenarios.

Compared with existing methods, SOMDE is more similar to those statistical models. This is related to our choice of the Gaussian process as the significance test model. If the self-organizing map is considered as the frontend and Gaussian process as the backend of SOMDE, the backend statistical model can take any other forms such as the marked point process or GLSM.

SOMDE can be easily extended to 3D spatial sequencing data in the future, while the computational complexity of other methods will further increase. Further studies are needed in the future to introduce more effective statistical models and explore ways to automatically select the appropriate resolution of SOM through data characteristics.

Acknowledgements

The authors wish to thank Prof. Guocheng Yuan, Dr. Qian Zhu and Dr. Ruben Dries for helpful suggestions and comments.

Funding

This work has been supported by the NSFC Projects (61721003) and National Key R&D Program of China (2018YFC0910401).

Conflict of Interest: none declared.

References

Dries,R. *et al.* (2019) Giotto, a pipeline for integrative analysis and visualization of single-cell spatial transcriptomic data Bioinformatics, *bioRxiv,* **2019**, 701680.

Edsgärd,D. *et al.* (2018) Identification of spatial expression trends in single-cell gene expression data. *Nat. Methods*, **15**, 339–342.

Eng,C.-H.L. *et al.* (2019) Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature*, **568**, 235–239.

Fan,Z. *et al.* (2019) SpatialDB: a database for spatially resolved transcriptomes. *Nucleic Acids Res.*, **48**, 2019.

Kohonen,T. (1982) Self-organized formation of topologically correct feature maps. *Biological cybernetics*, **43**(1), 59-69.

Lein,E.S. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.

Ling,K.H. *et al.* (2011) Spatiotemporal regulation of multiple overlapping sense and novel natural antisense transcripts at the Nrgn and Camk2n1 gene loci during mouse cerebral corticogenesis. *Cerebral cortex* **21.3** 683-697.

Moffitt,J.R. *et al.* (2018) Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science*, **362**, no.6416.

Rodriques,S.G. *et al.* (2019) Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*,**363**, 1463-1467.

Smith,C.M. *et al.* (2019) The mouse Gene Expression Database (GXD): 2019 update. *Nucleic Acids Res.*, **47** (D1), D774–D779.

Ståhl,P.L. *et al.* (2016) Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, **353**, 78–82.

Sun,S. *et al.* (2020) Statistical Analysis of Spatial Expression Patterns for Spatially Resolved Transcriptomic Studies. *Nat. Methods*, **17**, 193–200.

Svensson,V. *et al.* (2018) SpatialDE: identification of spatially variable genes. *Nat. Methods*, **15**, 343–346.

Uriarte E. *et al.* (2005) Topology preservation in som. *Int J Appl Math Comput Sci* **1** (1),19–22

Wittek,P. *et al.* (2017) Somoclu: An Efficient Parallel Library for Self-Organizing Maps. *Journal of Statistical Software*, **78**, 1–21.

Zhang,K. *et al.* (2018) Identification of Spatially Variable Genes with Graph Cuts. *bioRxiv*, **2018**, 491472.