RESEARCH

SMASH: Scalable Method for Analyzing Spatial Heterogeneity of genes in spatial transcriptomics data

Souvik Seal,[1*], Benjamin G. Bitler,[2] and Debashis Ghosh[3]

[*]Correspondence: sealso@musc.edu
[1]Department of Public Health Sciences, School of Medicine, Medical University of South Carolina, Charleston, USA
Full list of author information is available at the end of the article

Abstract

In high-throughput spatial transcriptomics (ST) studies, it is of great interest to identify the genes whose level of expression in a tissue covaries with the spatial location of cells/spots. Such genes, also known as spatially variable genes (SVGs), can provide crucial biological insights into both structural and functional characteristics of complex tissues. Existing methods for detecting SVGs either suffer from huge computational demand or significantly lack statistical power. We propose a non-parametric method termed SMASH that achieves a balance between the above two problems. We compare SMASH with other existing methods in varying simulation scenarios demonstrating its superior statistical power and robustness. We apply the method to four ST datasets from different platforms revealing interesting biological insights.

Keywords: Spatial transcriptomics; Single cell imaging datasets; Spatially variable genes; Non-paramteric method; Covariance Modelling; Slide-seq V2; MERFISH; 10X Visium

Background

Spatial transcriptomics (ST) performs high-throughput measurement of transcriptomes in complex biological tissues at single-cell or subcellular resolution, preserving spatial information [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. In the past decade, the rapid development of ST technologies has facilitated exciting discoveries in different domains, including neuroscience [11, 12, 13] and cancer research [14, 15, 16]. Early ST technologies, such as smFISH [17], seqFISH [18] and MERFISH [19], operate at relatively low spatial resolution, whereas recent technologies, such as Slide-seq [20], Slide-seq V2 [21], 10X Visium [22] and HDST [23], have enabled transcriptome-wide profiling at a much finer spatial resolution on multiple thousands of locations. Because of such huge spatial profiles, deriving biological insights from these datasets not only poses a plethora of statistical challenges but also demands maximal computational efficiency [24].

A critical step in the analysis of ST datasets is to identify the genes whose level of expression co-varies with the spatial locations across the tissue. These genes, often referred to as spatially variable genes (SVGs), can be used in downstream analyses, such as identifying potential markers for biological processes and defining areas in the tissue that dictate cellular differentiation and function [25, 26, 27, 28, 29]. For example, Wang et al. (2020) [30] analyzed an ST dataset on the tumor microenvironment (TME) of three tissue sections from a prostate cancer subject

[31]. In every tissue section, a unique set of spatially variable metabolic genes were identified, which they argued, can be used to guide targeted tissue-specific therapy. A simplistic approach for detecting SVGs could be to identify spatially located layers or cell types (if any) based on either a priori biological knowledge or using popular softwares, such as RCTD [32] and Seurat [33], with the transcriptional profiles, and then check which genes are enriched more in a particular spatial layer or cell type. However, such an approach would only work if the cell types are spatially well-separated, and always be sensitive to the quality of the cell type-identification step [34]. In recent years, more sophisticated methods have been developed to identify SVGs, a systematic overview of some of which can be found in Li et al. (2021) [35]. The methods can be broadly classified into two types: a) methods based on statistical modeling and b) methods based on graphical modeling. The methods of type (a) are based on parametric or non-parametric statistical model assumptions, whereas the methods of type (b) are usually model-free and based on graphical networks, constructed directly from the spatial locations or using simplified spatial grids. Some of the notable methods of type (a) include Trendsceek [36], SpatialDE [34], SPARK [37], and SPARK-X [38]. Some of the methods of type (b) are HMRF [39], MERINGUE [40], SpaGCN [41] and SpaGene [42]. Being model-based, type (a) methods allow adjusting for covariates such as cell type, whereas type (b) methods claim to be more flexible in the sense of being model-free.

The statistical power of both types of methods greatly varies based on gene expression patterns and the spatial structure of an ST dataset. The methods encounter different levels of computational complexity based on two quantities, N and K, corresponding to the numbers of cells/spots and genes, respectively. SpatialDE [34] is one of the earliest methods of type (a). It employs a Gaussian process (GP) regression model [43] with kernel-based covariance matrices [44] of multiple types, such as linear, Gaussian, and cosine, computed using the distance between the spatial coordinates of the cells. The model decomposes the total variability of a gene expression into two components, spatial and error variance. A significantly large value of the spatial variance would imply that the gene is spatially variable. Borrowing an efficient estimation algorithm from the statistical genetics literature [45], SpatialDE manages to estimate the variance components with a reasonable degree of computational efficiency, requiring $O(N^3 + NK^2)$ floating point operations (FLOPS). A newer method named SPARK [37] extends the framework of SpatialDE by considering a generalized linear spatial model (GLSM) [46] with a Poisson distribution, arguing to be better suited for modeling the raw count data usually obtained from most of the ST platforms. However, the penalized quasi-likelihood (PQL) approach [47] used for parameter estimation in SPARK is extremely computationally demanding with a complexity of $O(N^3K)$, making it unusable for a transcriptome-wide analysis when N is moderately large ($N > 3,000$). To this end, a non-parametric highly scalable method named SPARK-X [38] has been recently developed requiring just linear complexity w.r.t. N. It is based on the robust covariance testing framework [48] that compares the linear kernel-based covariance matrices of the gene expression and the spatial coordinates. However, using a linear kernel makes SPARK-X equivalent to fitting a multiple linear regression model [49] with the gene expression as the dependent variable and the spatial coordinates (or, some transformation of

these) as the predictors and testing if the fixed effect coefficients differ from zero. Thus, it is only capable of detecting spatial dependencies or patterns that manifest linearly in the mean or expected value of the gene expression, also known as first-order dependencies, and drastically loses power in complex scenarios as to be shown later. Zhu et al. (2021) [38] have partially acknowledged this issue with their main focus being computational scalability.

On the other side, a popular method of type (b), MERINGUE [40] considers spatial autocorrelation and cross-correlation based on spatial neighborhood graphs to identify SVGs. Improving hugely on the complexity of MERINGUE, another model-free method named SpaGene [42] has been recently developed. It constructs a spatial network between cells/spots using the k-nearest neighbors approach, and then for each gene, extracts the subnetwork whose nodes have high gene expression. Then, it compares the observed degree distribution of the subnetwork to a distribution from a fully connected network using the earth mover's distance [50]. It considers a permutation test [51] to obtain the p-value for every gene. SpaGene is highly comparable to SPARK-X w.r.t. computational complexity and thus applicable to ST datasets with large N. However, the method is harder to interpret than the methods of type (a), can not readily accommodate additional covariates, and also lacks power in various scenarios (see Simulations section).

We propose a non-parametric method, named SMASH, which achieves superior statistical power than both SPARK-X and SpaGene, while remaining computationally tractable. It augments the idea of SPARK-X in its use of the robust covariance testing framework [48] coupled with more general kernel-based spatial covariance matrices. With a computational complexity quadratic in N, SMASH sacrifices some degree of computational efficiency in favor of significantly higher detection power than both SPARK-X and SpaGene. However, it is worth highlighting that SMASH is notably faster than other type (a) methods, such as SpatialDE and SPARK, and can thus be thought of as a balanced alternative, fusing high detection power with a moderate degree of scalability. In varying simulation scenarios, we demonstrate that SMASH achieves highly consistent and superior performance as compared to the methods SPARK-X and SpaGene. Finally, our analysis of four large ST datasets from platforms like SlideSeq V2, 10X Visium, and MERFISH using these three methods not only reveals exciting biological insights but also demonstrates SMASH's capability of detecting SVGs that will be otherwise missed by either of the other two methods. A Python-based software implementation of SMASH is available at, https://github.com/sealx017/SMASH-package, which returns the lists of SVGs detected by both SMASH and SPARK-X, allowing users to investigate the overlap between them.

Results

Simulations

We evaluated the performance of SMASH, SPARK, and SpaGene in three different simulation setups. We omitted SpatialDE and SPARK from the power comparison for two reasons: a) high computational requirements and b) these two methods have already been greatly studied in previous works [38, 42]. In simulation setup (1), we followed the procedure described in the SPARK-X manuscript [38]. In setups (2) and

(3), we considered the Gaussian process (GP)-based spatial regression model from the SpatialDE manuscript [34], respectively with the Gaussian and cosine kernel-based covariance functions (see Equation (2)). In all the setups, three values of the number of cells (N) were considered, $N = 1000, 5000$, and 10,000. The spatial coordinates of the cells were simulated first and then the expression levels of K (500 or 1000) genes with varying levels of dependence. In setup (1), the expression levels were simulated using a negative binomial distribution, while in setups (2) and (3), those are simulated using a multivariate normal distribution. In all the setups, distinct spatial patterns were ensured to be present in the expression levels. Further details regarding the simulation setups are provided at the end of the Methods section. Figures 1, 2, and 3 respectively correspond to the three simulation setups, in which we display the simulated spatial patterns and the statistical power of the three methods for different parameter combinations.

In simulation setup (1), SMASH, and SPARK-X performed much better than SpaGene for all four spatial patterns, namely streak, reverse streak, hotspot, and reverse hotspot (Figure 1). SpaGene was particularly poor for the patterns: streak and hotspot. The power of SMASH and SPARK-X steadily increased as N and the fold-change parameter increased. Note that a fold value of 1 stood for no spatial association while a larger value indicated higher spatial association. This particular simulation setup favored SPARK-X in the sense that the spatial variability of the expression was of the first order, manifesting entirely through the mean or expectation. Even then SMASH could achieve similar power.

In simulation setups (2) and (3), the spatial variability of the expression was of higher order, manifesting through the covariance. In setup (2), which had the Gaussian covariance function, SMASH performed the best followed by SPARK-X and then SpaGene in most cases. SMASH performed the best in setup (3) as well. However, SpaGene achieved better power than SPARK-X here. SPARK-X had almost zero power in many of the cases, especially when the period p was small ($p = 0.5$, 1), demonstrating its lack of robustness under complicated spatial dependency structures.

We compared the run-time of the methods in the simulation setup (2) for varying numbers of cells, $N = 1000, 5000$, and 10000 (Table 1). Since the computational complexity of the algorithms mainly differ w.r.t. N and not the number of genes K, we kept $K = 1000$. We noticed that the run-time of SMASH expectedly increased in an almost squared order w.r.t. N. SPARK-X and SpaGene were both extremely fast for just having linear complexity w.r.t. N. We also added SpatialDE to this comparison just to show how computationally intensive it can be to fit a fully parametric model in such a context. We omitted SPARK entirely as it is much slower than even SpatialDE with a computational complexity of $O(N^3K)$.

Application to real data

We applied the methods, SMASH, SPARK-X, and SpaGene to four datasets: 1) mouse cerebellum data collected using Slide-seq V2 [21], 2) human dorsolateral prefrontal cortex (DLPFC) data collected using 10X Visium [12], 3) small cell ovarian carcinoma of the ovary hypercalcemic type (SCCOHT) data collected using 10X Visium [12], and 4) mouse hypothalamus data collected using MERFISH [52]. The datasets have varying numbers of genes and spots/cells.

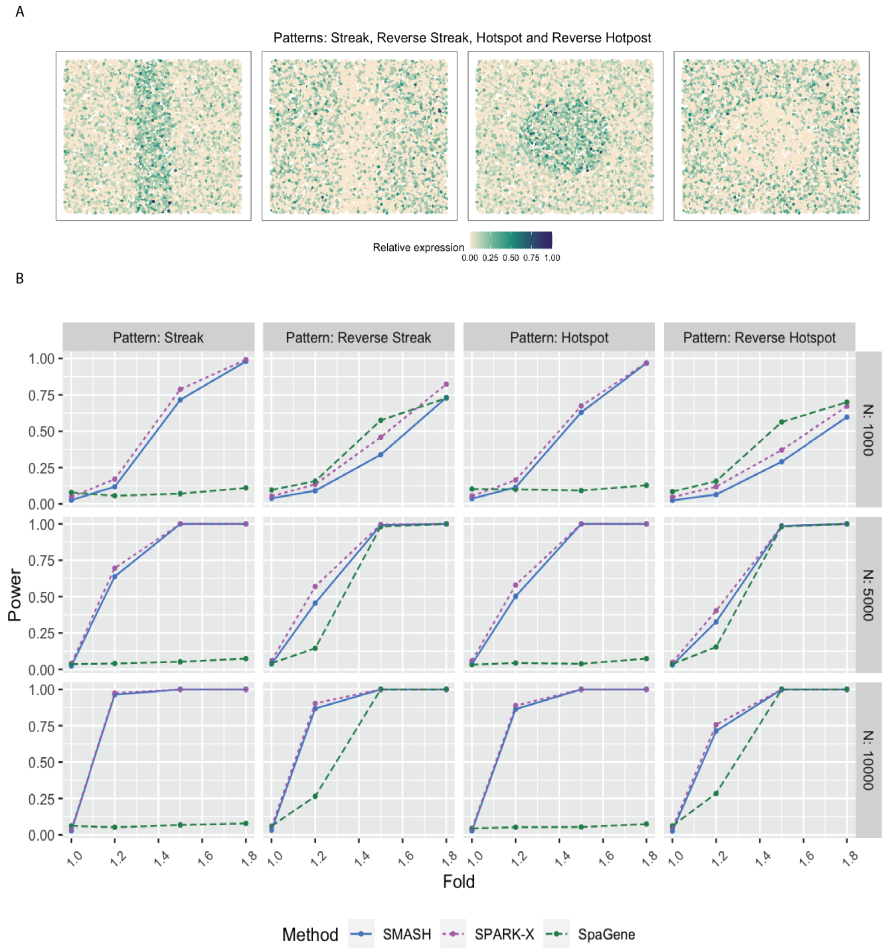

Figure 1 Simulation following the SPARK-X manuscript. A Four spatial expression patterns that the genes were assumed to follow. **B** Statistical power plots of the three methods, SMASH, SPARK-X, and SpaGene under varying values of N and fold-size, for $K = 500$ genes at a level of $\alpha = 0.05$. The results were averaged over five replications.

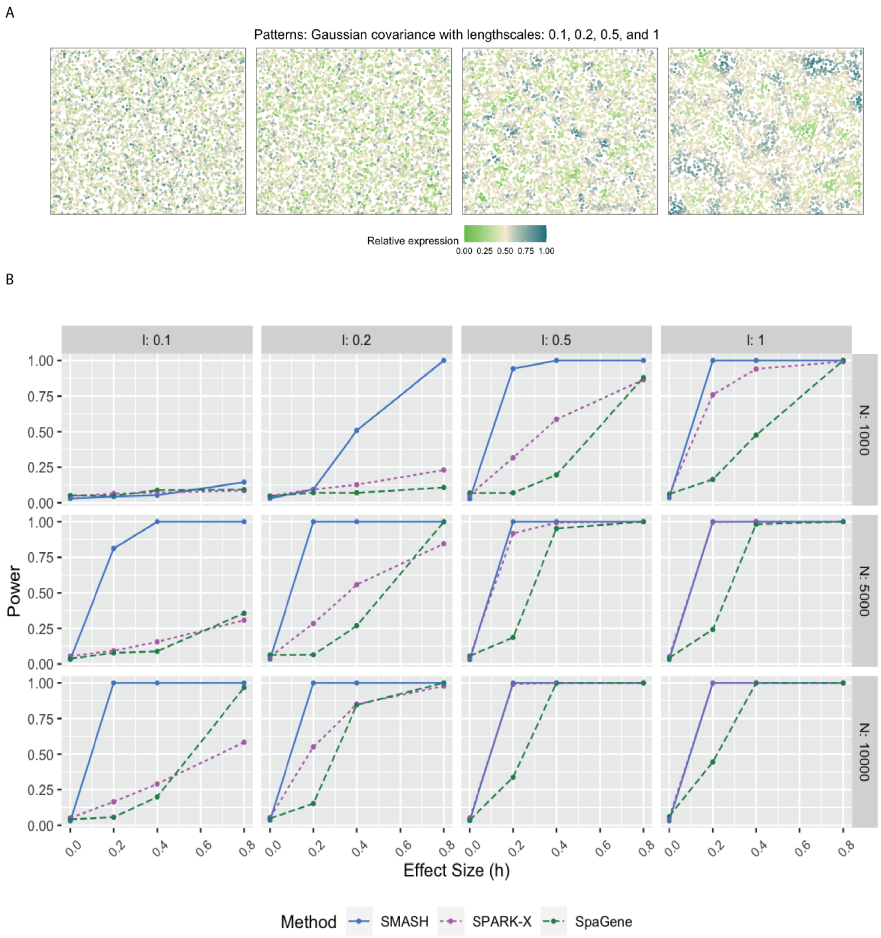

Figure 2 Simulation using Gaussian process-based regression model with the Gaussian covariance. A) Four spatial expression patterns that were generated using Gaussian covariance matrices with four different values of the lengthscale l. B) Statistical power plots of the three methods under varying values of N and effect-size (h) for $K = 1000$ genes at a level of $\alpha = 0.05$. The results were averaged over five replications.

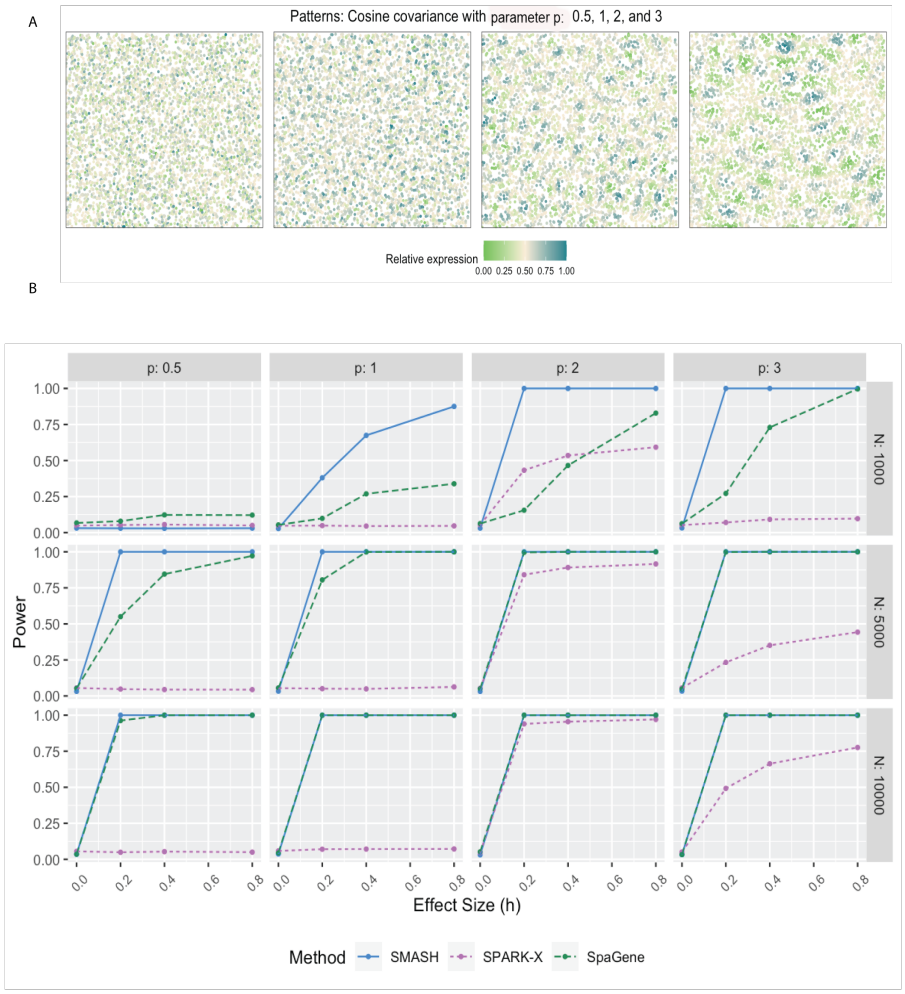

Figure 3 Simulation using Gaussian process-based regression model with the cosine covariance. A) Four spatial expression patterns that were generated using cosine covariance matrices with four different values of the period p. B) Statistical power plots of the three methods under varying values of N and effect-size (h) for $K = 1000$ genes at a level of $\alpha = 0.05$. The results were averaged over five replications.

Table 1 Computational complexity and run-time comparison. The table lists the theoretical complexity and run-time (in seconds) of the four methods, SMASH, SPARK-X, SpaGene and SpatialDE in a simulation setup with $K = 1000$ genes and varying number of cells N. The number of spatial co-ordinates d was equal to 2. * SpaGene constructs multiple kNN graphs and performs permutation tests. We are only listing the complexity of the KNN algorithm.

Method	Complexity	N = 1000	N = 5000	N = 10000
SMASH	$O(N^2K)$	0.9	18.5	97.5
SPARK-X	$O(NKd^2)$	0.72	4.1	4.3
SpaGene	$O(NKd)^*$	0.2	1.1	2.1
SpatialDE	$O(N^3 + NK^2)$	24.5	245	971.7

Mouse cerebellum by Slide-seqV2

The mouse cerebellum data [21] has 20,117 genes and 11,626 spots. We restricted our focus to the 7,653 genes that express in more than 1% of the spots. The mouse cerebellum is made of four spatial layers, white matter layer (WML), granule layer (GL), Purkinje layer (PL), and molecular layer (ML) [53]. These layers consist of different types of cells. For example, WML contains oligodendrocytes, GL contains granule cells, PL contains Purkinje neurons and Bergmann gila, and ML contains intra-neurons MLI. These cell types can be inferred based on just the transcriptional profiles using cell clustering software like RCTD [32]. We display the inferred cell types overlayed on the spatial locations in Figure 4. Out of the 7,653 genes, SMASH identified 1173 genes to be spatially variable (adjusted p-value: $p_{\text{adjust}} < 0.05$). SPARK-X and SpaGene respectively detected 608 and 518 genes, and the overlaps between the detected SVGs by the three methods are displayed in a Venn diagram (Figure 4). We noted that SPARK-X and SpaGene had many of the SVGs uncommon. SMASH, on the other hand, could identify almost all the detected genes by those two methods, especially SPARK-X, while detecting an additional 363 SVGs.

Next, we performed two types of enrichment analysis. First, we compared the performance of the methods in different layers by computing their enrichment scores (ES) following Liu et al. (2022) [42]. It is based on the expectation that the genes which abundantly express themselves in the four spatial layers, should be identified and ranked top by the methods. In that regard, we noticed that SPARK-X performed poorly in the PL, whereas SpaGene performed poorly in the WML. SMASH, on the other hand, consistently achieved similar or better performance compared to the other two methods in all four layers. Next, we individually performed functional enrichment analysis on the following four sets of SVGs: a) the common genes identified by all three methods, b) the genes identified by SMASH and SpaGene but not by SPARK-X, c) the genes identified by SMASH and SPARK-X but not by SpaGene, and d) the genes identified only by SMASH. The expression pattern of three representative genes of the enriched pathways for each of these four sets of genes, are shown in Figure 5. For set (a), top enriched Gene Ontology (GO) terms, such as GO: 0098916 (anterograde trans-synaptic signaling), GO: 0007268 (chemical synaptic transmission), and GO: 0099536 (synaptic signaling), were broadly associated with synaptic regulation. The protein-coding genes Fam107a, Ppp3ca, and Calm1 appeared in these top pathways. Fam107a seems to express in the PL, whereas the other two express in the GL (Figure 5). For set (b), the top GO terms including GO: 0006873 (intracellular monoatomic ion homeostasis), GO: 0030003 (intracellular monoatomic cation homeostasis), and GO: 0098771 (inorganic ion homeostasis) were associated with ion homeostasis. The representative genes Atp1a3 and Thy1

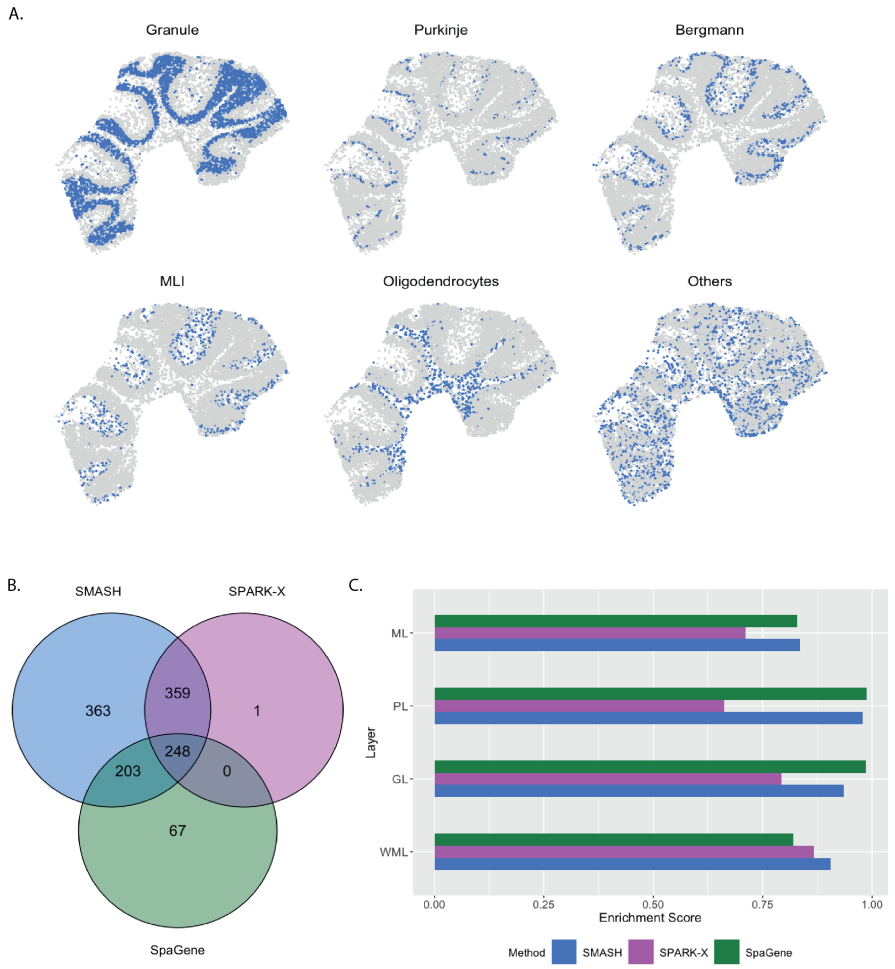

Figure 4 Analysis of mouse cerebellum data. A) Location of the major cell types corresponding to the four spatial layers of the mouse cerebellum. B) Overlap between the detected SVGs by the three methods. C) Enrichment scores of the methods in the four spatial layers.

express in the PL while Calm3 expresses in the GL. For set (c), the top pathways including GO: 0006811 (monoatomic ion transport), GO: 0006812 (monoatomic cation transport), and GO: 0098655 (monoatomic cation transmembrane transport) were associated with ion transportation. The representative genes Pllp and Efnb3 express in the WML, whereas Cox7a2 expresses roughly in the GL. For set (d), the top enriched GO terms, such as GO: 0044057 (regulation of system process) and GO: 0050877 (nervous system process), were associated with regulating different types of system processes. The representative genes Gls, Tmem36a, and Coro2b roughly express in the GL.

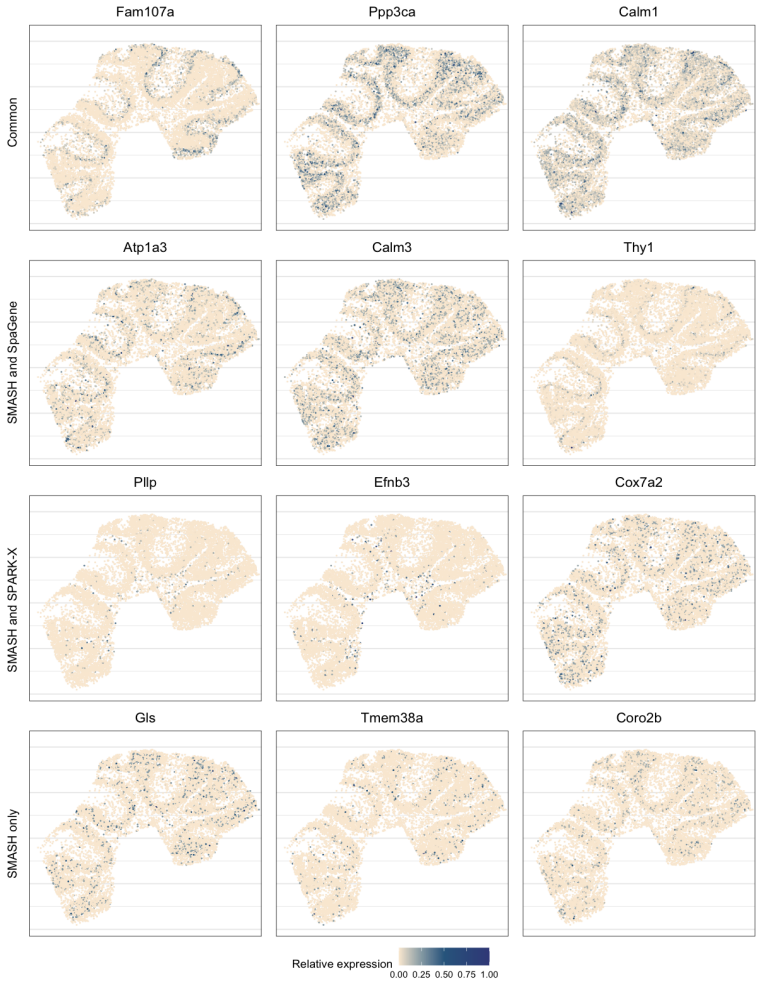

Figure 5 Expression patterns in mouse cerebellum data. Three representative genes from the detected pathways for the four sets of genes: a) the common genes identified by all three methods, b) the genes identified by SMASH and SpaGene but not by SPARK-X, c) the genes identified by SMASH and SPARK-X but not by SpaGene, and d) the genes identified only by SMASH.

Human DLPFC by 10X Visium

The human dorsolateral prefrontal cortex (DLPFC) data [12] has 33,538 and 3,639 spots. We focused on the 13,783 genes which express in more than 1% of the spots. Every spot belongs to one of the six manually labeled cortical layers or the white

matter layer (WML) (Figure 6). SMASH and SPARK-X identified 10,871 and 10,416 SVGs respectively ($p_{\text{adjust}} < 0.05$), whereas SpaGene identified only 2379. The overlaps between the detected SVGs by the three methods are displayed in a Venn diagram (Figure 6). We noted that almost all the genes detected by SpaGene were also detected by both SMASH and SPARK-X. SMASH and SPARK-X detected a lot of additional SVGs. We performed functional enrichment analysis on the two sets of detected genes: a) the common genes identified by all three methods and b) the genes identified only by SMASH and SPARK-X but not by SpaGene. For set (a), top enriched GO terms, such as GO: 0099537 (trans-synaptic signaling) and GO: 0099177 (regulation of trans-synaptic signaling), were associated with synaptic signaling. For set (b), top enriched GO terms like GO: 0006397 (mRNA processing) and GO: 0000375 (RNA splicing, via transesterification reactions), were associated with RNA processing. The expression of three representative genes from the set (b) are displayed in Figure 6. There seemed to be a gradient spatial pattern of expression for all three genes which SpaGene failed to detect. Similar to the previous section, we computed the enrichment score (ES) of every method in the seven manually labeled spatial layers. From Figure 6, we noticed that SpaGene performed poorly in terms of ES, especially in Layers 1 and 6. We also performed an additional check as follows. There are three cortical-layer associated SVGs, MOBP, SNAP25, and PCP4, and three blood and immune-related SVGs, HBB, IGKC, and NPY, known to be spatially variable from previous studies [12]. We checked how many of these genes appeared in the lists of the top thousand SVGs (in terms of p_{adjust}) by the three methods. SMASH and SpaGene respectively ranked five and six of these SVGs, whereas SPARK-X ranked only two cortical-layer associated genes.

SCCOHT by 10X Visium

The small cell carcinoma of the ovary hypercalcemic type (SCCOHT) data [54] has 15,229 genes and 2071 cells. We restricted our focus to the 12,001 genes that express in more than 5% of the cells. Sanders et al. (2022) [54] grouped the cells into twelve clusters based on the expression profile of a selected few genes, using Seurat [33], which we display in Figure 7. SMASH, SPARK-X, and SpaGene respectively detected 9361, 6564, and 6899 SVGs ($p_{\text{adjust}} < 0.05$). The overlaps between the detected SVGs by the three methods are displayed in a Venn diagram (Figure 7). SMASH could detect most of the SVGs identified by at least one of the other two methods and an additional 1634 genes. Similar to the analysis of the Mouse cerebellum data, we checked if the methods could identify the top genes that show enriched expression in the twelve spatially well-separated clusters found by Sanders et al. (2022). We computed the enrichment scores (ES) of the methods for each of the clusters (Figure 7). SMASH achieved consistently higher ES for all the clusters while SpaGene was the second best in most cases. Additionally, in Figure 8, we show the expression of three chosen genes from each of the following four sets of SVGs, a) the common genes identified by all three methods, b) the genes identified by SMASH and SpaGene but not by SPARK-X, c) the genes identified by SMASH and SPARK-X but not by SpaGene, and d) the genes identified only by SMASH. We also checked the clinical relevance of these genes in the existing literature. For example, CITED4, which was detected to be an SVG by all three methods, has been

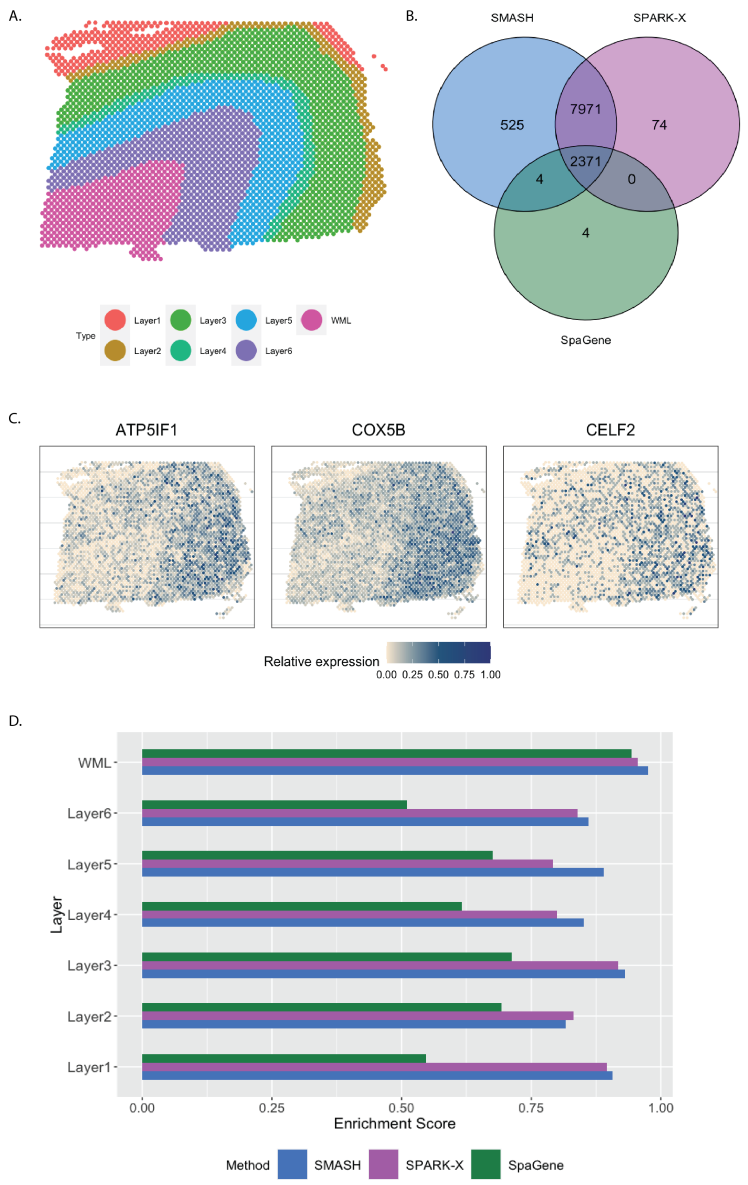

Figure 6 Analysis of human DLPFC data. A) Manually labeled cortical layers (layers 1-6) and white matter layer (WML). B) Overlap between the detected SVGs by the three methods. C) Expression of three representative genes identified only by SMASH and SPARK-X. D) Enrichment scores of the methods in different layers.

found to be associated with lung adenocarcinoma [55]. From the set (b), ELF4A1 has been found to be associated with gastric cancer [56]. EZH2, from the set (c), is a well-known marker for being associated with the development and progression of different types of cancer [57, 58]. Sanders et al. (2022) [54] also found the expression of EZH2 to be highly variable across their identified spatial clusters. Finally, from the set (d), SEMA4F has been found to be associated with endometrial cancer [59].

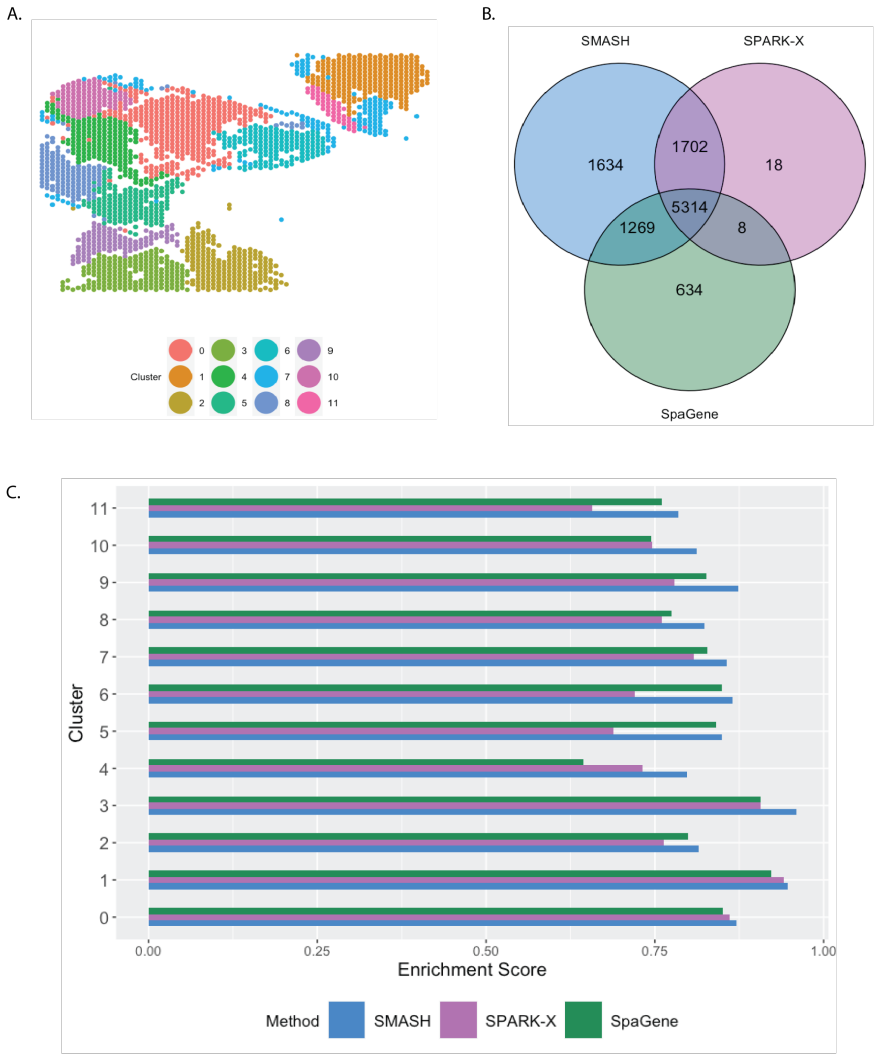

Figure 7 Analysis of SCCOHT data. A) Pre-identified clusters of cells using Seurat. B) Overlap between the detected SVGs by the three methods. C) Enrichment scores of the methods in different clusters.

Mouse hypothalamus by MERFISH

The mouse hypothalamus data [52] has 161 genes and 5665 cells. 156 genes are pre-selected markers for different cell types and can thus be expected to be highly variable, whereas the other five are control genes. The cell types, such as endothelial, ependymal, and inhibitory, can be identified based on the transcriptional profiles of the markers. The spatial organizations of a few major cell types are shown in

Figure 8 Expression patterns in SCCOHT data. Three representative genes from the four sets of SVGs: a) the common genes identified by all three methods, b) the genes identified by SMASH and SpaGene but not by SPARK-X, c) the genes identified by SMASH and SPARK-X but not by SpaGene, and d) the genes identified only by SMASH.

Figure 9. SMASH was able to detect 139 genes, whereas SPARK-X and SpaGene detected 127 and 124 genes, respectively ($p_{\text{adjust}} < 0.01$). The overlaps between the SVGs detected by the three methods are shown in Figure 9. SMASH identified all the SVGs SPARK-X could detect, while SpaGene identified one additional SVG. It should be highlighted that all the methods assigned the five control genes to not be spatially variable. We display the expression of two representative genes from three sets of genes, a) the genes identified only by SMASH and SpaGene, b) the genes identified only by SMASH and SPARK-X, and c) the genes identified only by SMASH. We did not focus on the common genes because they have been extensively studied in earlier literature, such as the work of Liu. et al. (2022)[42]. The genes Npy1r and Cplx3 belonged to set (a), and are known to be enriched in inhibitory and excitatory neurons [60, 61]. Rxfp1 and Ntsr1 belonged to set b). Even though both genes are known to express in inhibitory and excitatory neurons, Rxfp1 seems to express in ependymal cells as well. Galr2 and Crhr1 are two genes from set c) which express in multiple cell types including inhibitory cells and astrocytes.

Discussion

We have proposed a novel non-parametric method SMASH for detecting spatially variable genes (SVGs) in the context of large-scale spatial transcriptomics (ST) datasets. In comparison to existing scalable approaches, SMASH achieves superior power in both complex simulation scenarios and real data analyses while remaining computationally tractable.

Recently developed spatial transcriptomics platforms produce high-dimensional datasets [20, 21, 23] in terms of the number of cells and the number of genes. In such large datasets, fully parametric approaches for detecting SVGs, such as SpatialDE [34] and SPARK [37], albeit statistically powerful, become unusable for their high computational demand. Computationally efficient alternative non-parametric approaches, such as SPARK-X [38] and SpaGene [35], on the other hand, can often turn out to be significantly less powerful. In our method SMASH, we strive to find a balance between these two issues, meaning that we achieve higher statistical power while achieving a moderate degree of scalability. We augment the kernel-based covariance testing framework [48], used before in SPARK-X, by accounting for more complex spatial dependencies.

In three different simulation setups, one following the SPARK-X manuscript [38] and the other two following the framework of SpatialDE [34], we evaluated the performance of SMASH, along with two other methods: SPARK-X and SpaGene, in terms of type 1 error and power. SMASH achieved consistently similar or better power than the other two methods in all the simulation setups for all combinations of the varying parameters. In contrast, both SPARK-X and SpaGene behaved unpredictably, achieving almost zero detection power in many of the cases. It demonstrated their lack of statistical robustness and failure to capture complicated structures of spatial dependency in the gene expression. In the run-time comparison of the methods, we showed that SMASH, although slower than SPARK-X and SpaGene, remained fairly tractable and was almost ten times faster than a fully parametric approach like SpatialDE. SMASH, SPARK-X, and SpaGene were then applied to four real datasets: 1) mouse cerebellum data collected using Slide-seq

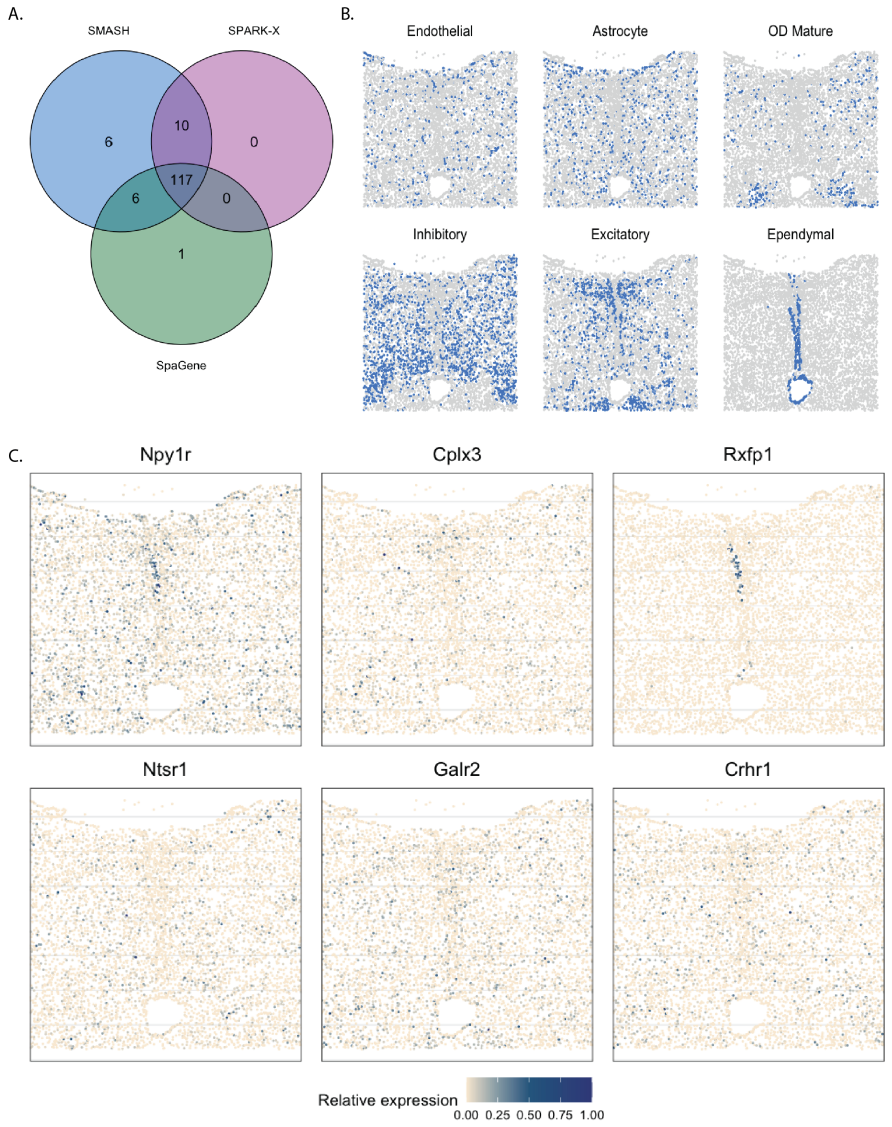

Figure 9 Analysis of mouse hypothalamus data. A) Overlap between the detected SVGs by the three methods. B) Spatial organization of a few major cell types. C) Expression of two representative genes from each of the three sets, a) the genes identified only by SMASH and SpaGene, b) the genes identified only by SMASH and SPARK-X, and c) the genes identified only by SMASH.

V2 [21], 2) human dorsolateral prefrontal cortex data collected using 10X Visium [12], 3) small cell ovarian carcinoma of the ovary hypercalcemic type data collected using 10X Visium [12], and 4) mouse hypothalamus data collected using MERFISH [52]. We compared the methods via a number of avenues: a) checking the overlap between the detected SVGs by the three methods, b) computing enrichment scores (ES) of the methods in different spatial layers or cell types identified based on the transcriptional profiles using popular softwares, such as RCTD [32] and Seurat [33], and c) investigating the functional enrichment of the genes that were detected by SMASH but remained undetected by at least one of the other two methods. For all the datasets, SMASH detected more SVGs than the other two methods, which included nearly all of the SVGs detected by SPARK-X. SMASH could also detect most of the SVGs that were identified by SpaGene but not by SPARK-X. For example, in data (1), from the 7,653 genes after quality control, SMASH identified 1173 SVGs which included 607 out of the 608 SVGs SPARK-X could detect. Out of the 518 SVGs detected by SpaGene, only 248 were also detected by SPARK-X, while SMASH detected 451 of them. It is important to highlight that SMASH produced calibrated p-values in the null simulations from all of these datasets, lending credibility to these higher numbers of detected SVGs. In the same dataset, SMASH achieved a higher enrichment score (ES) than the other two methods in different pre-identified spatially separated layers or cell types of the mouse cerebellum. A higher ES meant better capability to identify the genes that showed highly variable expression in a particular spatial layer compared to the rest. In the other datasets as well, SMASH consistently achieved better ES in different spatially localized cell types. We also studied the functional properties and clinical significance of the identified SVGs. For example, in data (3), the gene EZH2 was detected to be spatially variable by SMASH and SPARK-X. EZH2 is a known marker for the progression of multiple types of cancers [57, 58]. It has also previously been found to be highly variable in some particular spatially localized cell clusters by the group of researchers who performed the original experiment [54].

In all the methods we have discussed, including SMASH, the biology of a single tissue section from a single subject is explored at a time. It means that if we either have multiple tissue sections from the same subject or from multiple subjects, the methods will have to identify SVGs individually, disregarding the potential of harnessing shared information between and across the subjects. Thus, we would like to extend SMASH in a hierarchical fashion for jointly analyzing more than one tissue section or subject in the future. One more important functionality that we would like to incorporate would be the ability to classify the genes based on their similarity of spatial expression patterns. For example, SpatialDE [34] considers a hierarchical Bayesian mixture model approach that suffers from extremely high computational demand. SpaGene [42] considers a non-negative matrix factorization [62] of the expression data to identify similarly expressed genes. This approach, although computationally feasible, does not take into account the spatial locations directly and can thus be suboptimal in capturing truly spatial patterns. In the future, we would like to study this problem with a deeper focus and pursue methodological development in this area. Finally, we would like to explore the possibility of using SMASH in the context of multiplex immunohistochemistry (mIHC) datasets [63, 64] where the goal is to identify spatially variable cell types and their interaction.

Methods

We briefly discuss some of the existing methods such as SpatialDE [34], SPARK [37], SPARK-X [38], and SpaGene [35], and then present the proposed method SMASH. Note that we did not compare SMASH to either SpatialDE or SPARK in our Results section except for the time comparison, primarily due to their high computational demand and the fact that these have already been studied in great detail in earlier works. However, it is crucial to talk about their modeling frameworks for the sake of understanding the other methods better. Let us introduce a few relevant notations. Suppose there is a single subject (image) with N cells/spots and the expression profile of K genes is observed in the cells. For the i-th cell, let s_i denote its location i.e., a vector of spatial (two or three-dimensional) coordinates, and y_{ki} denote the expression of the k-th gene in the cell. Let us also define, $y_k = (y_{k1}, \ldots, y_{kN})^T$ and $S = (s_1, \ldots, s_N)^T$. For the sake of simplicity, we are assuming that there are no additional covariates but in all the methods, except SpaGene, covariates can be readily incorporated.

A brief overview of existing methods

SpatialDE

SpatialDE uses a Gaussian process (GP)-based spatial regression model [43, 65]. It assumes that at every location $s \in \mathcal{S} \subseteq \mathbb{R}^3$, the expression of the k-th gene is a process $y_k(s)$ of the following form,

$$y_k(s) = \mu_k + w_k(s) + \epsilon_k(s); \quad w_k(s) \sim \text{GP}(0, \tau_k^2 C(s, s')); \quad \epsilon_k(s) \sim N(0, \sigma_k^2); \quad (1)$$

where $w_k(s)$ is a zero-centered GP with variance τ_k^2 and covariance function $C(s, s')$ for $s' \in \mathcal{S}$. A popular choice for covariance function is the Gaussian kernel-based covariance, $C(s, s') = \exp\left[-||s - s'||^2/2l^2\right]$, where $||.||$ denotes the Euclidean norm, and the hyperparameter l, known as the characteristic lengthscale [66], controls the rapidness at which the covariance decays as a function of the spatial distance. $\epsilon_k(s)$ is an independent process with variance σ_k^2. With the observed values of s and $y_k(s)$, the data likelihood corresponding to Equation (1) can then be written as,

$$y_k \sim N(\mu_k \mathbf{1}, \tau_k^2 \Sigma + \sigma_k^2 I)); \quad \Sigma = [[\Sigma_{ij}]]_{N \times N}; \quad \Sigma_{ij} = \exp\left[-\frac{||s_i - s_j||^2}{2l^2}\right]; \quad (2)$$

where $\mathbf{1}$ denotes the n-length vector of all 1's, I denotes the n-dimensional identity matrix and Σ denotes the corresponding Gaussian covariance matrix. The fixed effect μ_k accounts for the mean expression level and τ_k^2 accounts for the expression variance attributable to spatial effects. A large value of τ_k^2 should imply that the gene shows differential spatial expression. To formally test the hypothesis, $H_0 : \tau_k^2 = 0$ against $H_1 : \tau_k^2 > 0$, SpatialDE considers the likelihood ratio test (LRT) [67]. To estimate the model parameters under the full model, the log-likelihood corresponding to Equation (2) is optimized w.r.t. $(\mu_k, \tau_k^2, \sigma_k^2)$ using an efficient algorithm by Lippert et al. (2011) [45]. Ideally, it is desirable to optimize over the hyperparameter l as well but for the sake of computational feasibility, l is kept fixed at a few carefully chosen values. For every choice of Σ, to analyze all K genes, the efficient algorithm requires just one computationally demanding step with a complexity of $O(N^3)$,

instead of $O(N^3K)$ as incurred in naive algorithms. Along with the Gaussian covariance function, SpatialDE also considers linear and cosine covariance functions to construct Σ, and finally, combines all the LRT values corresponding to different choices of Σ for the inference. For a particular Σ, the computational complexity of SpatialDE is of $O(N^3 + N^2K)$.

Note that, SpatialDE essentially assumes that gene expression is a continuous random variable (r.v.) and follows a normal distribution. However, for most technologies, gene expression is obtained as count data. Normalizing the count data has been shown to be sub-optimal in several omics studies [68, 69]. Improving upon that, newer methods named SPARK [37] and SPARK-X [38] have been proposed which we discuss next.

SPARK and SPARK-X

SPARK [37] considers a generalized linear spatial model (GLSM) [46] with Poisson distribution, which is a generalization of Equation (1) with the following form,

$$y_k(s) \sim Poi(N(s)\lambda_k(s)); \quad \log(\lambda_k(s)) = \mu_k + w_k(s) + \epsilon_k(s); \tag{3}$$

For cell i, $\lambda_k(s_i)$ is an unknown Poisson rate parameter that represents the underlying gene expression. The processes $w_k(s)$ and $\epsilon_k(s)$ are defined as earlier and the variance parameters, τ_k^2 and σ_k^2 have a similar interpretation as earlier. To test $H_0 : \tau_k^2 = 0$, SPARK uses the score test [70]. Parameter estimation and inference are incredibly hard in GLSM which is why SPARK uses an approximate algorithm based on the penalized quasi-likelihood (PQL) approach [47, 71]. The approach has the computational complexity of $O(N^3)$ for every trait, or $O(N^3K)$ in total. Thus, it lacks severely in terms of scalability. The SPARK *R* package also has a Gaussian model option (instead of Poisson) which is equivalent to SpatialDE and has a similar computational requirement.

Improving upon SPARK's scalability, a recent non-parametric method named SPARK-X [38] has been proposed. The method is built on a simple intuition: if y_k is independent of S, the spatial distance between two locations i and j should be independent of the difference in gene expression between the two locations. It computes the expression covariance matrix, $E_k = y_k(y_k^T y_k)^{-1} y_k^T$ and the distance covariance matrix, $D = S(S^T S)^{-1} S^T$ and constructs the test statistic as, $T_k^{\text{SPARKX}} \equiv tr(E_k D)/N$ where $tr()$ denotes the trace operator. We assume y_k to be mean-standardized here for the sake of simplicity. Under the null hypothesis of no association, T_k asymptotically follows a weighted mixture of independent χ_1^2 distributions. The weights are the products of the ordered eigenvalues of the matrices, E_k, and D. This test falls under a general class of covariance tests [48], including the Hilbert-Schmidt independence criteria test [72] and the distance covariance statistic [73]. SPARK-X requires the computational complexity of just $O(Nd^2)$ for every gene, or $O(NKd^2)$ in total, where d is the dimension of the location-space $\mathcal{S}$, e.g., $d = 3$ if $\mathcal{S} = \mathbb{R}^3$. Linearity of the complexity w.r.t. N makes SPARK-X easily applicable to large-scale transcriptomics datasets. To capture more complex dependencies between y_k and S, SPARK-X also considers several element-wise non-linear transformations of S as $g(S)$, where g is a Gaussian or cosine transformation

with fixed scale or period parameters. Then, it repeats the above testing procedure replacing S with $g(S)$, and finally, combines all the corresponding p-values using Cauchy p-value combination rule [74].

However, the form of D corresponds to a linear covariance function [75]. It makes SPARK-X equivalent to performing a multiple linear regression of y_k on S or $g(S)$ and testing if the fixed effect parameters differ from zero. Thus, SPARK-X is only capable of detecting first-order spatial dependencies and as shown in the Results section, severely lacks power for higher-order dependencies. Refer to the supplementary information for more discussion on this topic.

SpaGene

A very recently developed method, SpaGene [42] is quite different from the rest of the methods in the sense of being model-free and based on graphs. The intuition behind the method is that the cells/spots with high gene expression are more likely to be spatially connected than random. It constructs the k-nearest neighbor (kNN) graph based on spatial locations. Then, for each gene, it extracts a subnetwork comprising only cells/spots with high expression from the kNN graph. SpaGene quantifies the connectivity of the subnetwork using the earth mover's distance (EMD) [50] between degree distributions of the subnetwork and a fully connected one. To generate the null distribution of the EMD for inference, a permutation test is considered. For further details, we refer the readers to [41].

Proposed method: *SMASH*

We consider a test statistic based on the kernel-based test of independence developed by Zhang et al. (2012) [48] as, $T_k^{\text{SMASH}} \equiv tr(E_k H(S))/N$, where E_k is defined as earlier and $H(S)$ is any $N \times N$ kernel-based covariance matrix based on the locations S. In this work, we consider $H(S)$ to have three forms: a) the Gaussian kernel-based covariance matrix, Σ defined in Equation (2), b) a cosine kernel-based covariance matrix of the form, $H(S) = [[\cos(2\pi||s_i - s_j||/p)]]_{N\times N}$, where parameter p is known as the period, and c) the linear kernel-based covariance matrix D considered in SPARK-X (Gaussian and cosine transformations of S: $g(S)$ are considered as well in constructing D). For the Gaussian covariance matrix, we consider ten data-driven fixed values of the lengthscale l [66], and for the cosine covariance matrix, we consider ten data-driven fixed values of the period p. For detailed discussions on the relevance of using these types of covariance matrices in the context of ST datasets, we refer the readers to [34] and [37].

Similar to the SPARK-X test statistic, T_k^{SPARKX}, the asymptotic null distribution of T_k^{SMASH} will be a weighted mixture of independent χ_1^2 distributions, where the weights are the products of the ordered eigenvalues of the matrices, E_k, and $H(S)$. However, $H(S)$ does not always have a form like D and thus, its eigenvalues can not be computed with the complexity of $O(Nd^2)$. Instead, it requires the complexity of $O(N^3)$, which becomes intractable as N increases. It has been shown that the asymptotic null distribution of T_k^{SMASH} can be well approximated by a gamma

distribution [48] as below,

$$T_k^{\text{SMASH}} \sim \Gamma(\theta_1, \theta_2); \quad \theta_1 = \frac{\mathbb{E}(T_k^{\text{SMASH}})^2}{\mathbb{V}(T_k^{\text{SMASH}})}; \quad \theta_2 = \frac{\mathbb{V}(T_k^{\text{SMASH}})}{\mathbb{E}(T_k^{\text{SMASH}})};$$
$$\mathbb{E}(T_k^{\text{SMASH}}) = \frac{1}{N^2} tr(E_k) tr(H(S)); \quad \mathbb{V}(T_k^{\text{SMASH}}) = \frac{2}{N^4} tr(E_k^2) tr(H(S)^2);$$

where $\mathbb{E}()$ and $\mathbb{V}()$ denote the expectation and variance, respectively. It is easy to verify that $tr(E_k) = tr(E_k^2) = 1$. Notice that we can now avoid any operation of complexity $O(N^3)$. Computation of $tr(H(S)^2)$ just requires the complexity of $O(N^2)$ using the property that $tr(AB) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} b_{ij}$ for two matrices, $A = [[a_{ij}]]_{N \times N}$ and $B = [[b_{ij}]]_{N \times N}$. Thus, for a particular choice of $H(S)$, to analyze all K genes, SMASH requires the complexity of $O(N^2 K)$. This computational complexity is higher than SPARK-X. But we are making that sacrifice to gain significantly more power, as shown in both simulation studies and real data analyses while still achieving a moderate degree of scalability. It is worth pointing out that even though SMASH is non-parametric and does not make any distributional assumptions, T_k^{SMASH} shares a close connection with the SpatialDE model under some additional assumptions (see the supplementary information).

As mentioned earlier, we consider multiple (say, R) choices for $H(S)$, to construct multiple test statistics: $T_{kr}^{\text{SMASH}}, r = 1, \ldots, R$. Finally, we combine the p-values corresponding to these test statistics using the minimum p-value combination rule [74] (see the supplementary information for more details). Note that we have assumed that y_k is mean-standardized and there are no additional covariates to be taken into account. In presence of covariates, we would regress the covariates out from the gene expression vector y_k, prior to performing the test, using a multiple linear regression model. To further elaborate, letting X be the corresponding matrix of covariates, we would compute the projection matrix $P_X = X(X^T X)^{-1} X^T$, and substitute the vector y_k with $y_k^* = [I - P_X] y_k$, in our proposed test statistic. A Python-based software implementation of SMASH is available at, https://github.com/sealx017/SMASH-package. The software returns the SVGs detected by both SMASH and SPARK-X giving users a chance to examine the overlap between the methods.

FDR control

In the real data analysis, we used Benjamini-Yekutieli [76] procedure to control the false discovery rate (FDR) at 0.05 (or, 0.01) for all the methods. In the Results section, p_{adjust} refers to the adjusted p-values. It was shown by Zhu et al. (2021) [38] that parametric methods like SpatialDE and SPARK often produce highly inflated p-values for most ST datasets, and hence need additional testing correction. To check if our p-values were inflated in the four real datasets, we randomly permuted the spatial locations of the cells/spots five times and then performed the tests using the three methods. Thus, we obtained the empirical null distribution of the p-values for each method which we displayed as quantile-quantile plots (see Figure S1 in the supplementary information). In all four cases, SMASH showed no sign of inflation with rather slightly conservative p-values which is expected since the minimum p-value combination rule used for combining the p-values in our method, is known to be conservative [77].

Enrichment scores

In the real data analysis, we computed the enrichment scores (ES) of the three methods following the procedure outlined in Liu et al. (2022) [42]. Cell clustering based on biological knowledge or using popular softwares, such as RCTD [32] and Seurat [33], with the transcriptional profiles, can often identify spatially localized layers or cell types. Therefore, marker genes in those spatially-restricted cell types should ideally be identified as SVGs. Suppose there are M cell types. For every cell type m, the gene set G_m is built from the top 50 markers based on the fold change between the expression in the cell type m compared to the others. The SVGs detected by the three methods are ranked from the most to the least significant. Finally, unweighted gene set enrichment analysis [78] is implemented to evaluate the enrichment of the gene sets, $G_m, m = 1, \ldots, M$, in the high ranking of the ranked SVG lists of the methods.

Simulation description

In simulation setup (1), we generated the spatial coordinates for varying numbers of cells, $N = 1000, 5000$, and 10,000 using a random point-pattern Poisson process [79]. The expression values of $K = 500$ genes in these cells were simulated based on a negative binomial distribution displaying one of the four spatial patterns: streak, reverse streak, hotspot, and reverse hotspot as shown in Figure 1. For each of the patterns, 80% of the spatial locations were assumed to be background locations, while the rest 20% were assumed to be part of the pattern. The difference between the mean expression of a gene on a background location and a patterned location was captured through a fold-change parameter. Several values of fold-change were considered where a value of 1 implied a null scenario i.e., no spatial pattern, and a high value implied a prominent spatial pattern. We refer to Zhu et al. (2021) [38] for more details.

For simulation setup (2), we considered the Gaussian process (GP)-based spatial regression model from SpatialDE [34]. The locations were simulated based on Uniform distribution, which were then used to construct Gaussian covariance matrices with varying lengthscale (l) parameters as in Equation (2). The expression levels of genes were independently and identically simulated from the multivariate normal distribution described in Equation (2) for different values of the variance parameters τ_k^2 and σ_k^2. We fixed the total variance, $\tau_k^2 + \sigma_k^2 = 1$, and varied the individual values as $\tau_k^2 = h$ and $\sigma_k^2 = 1 - h$, where "effect-size" h ranged from zero to larger values implyinh null to an increasingly stronger spatial pattern. In simulation setup (3), we followed setup (2) replacing the Gaussian covariance with the cosine covariance for varying values of the period parameter p. In all three setups, we compared SMASH, SPARK-X, and SpaGene in terms of type 1 error and power.

Additional Files

Additional file 1: Supplementary information. It includes additional mathematical discussions pertaining to the proposed method, and one figure displaying the empirical null p-values of the three methods, SMASH, SPARK-X, and SpaGene in the four real datasets in consideration.

Acknowledgements

We would like to thank Dr. Kristen Wells-Wrasman for her help with processing the SCCOHT dataset.

Funding

S.S. was supported in part by the Biostatistics Shared Resource, Hollings Cancer Center, Medical University of South Carolina (P30 CA138313).

Availability of data and materials

A Python-based software implementation of SMASH is available at, https://github.com/sealx017/SMASH-package. The package provides two detailed notebooks to perform the analysis on the mouse hypothalamus data by MERFISH and the human DLPFC data by 10X Visium (along with the datasets as compressed Python objects). Both the mouse cerebellum data by Slide-seqV2 and the human DLPFC data by 10X Visium are available in the *R* Bioconductor package: STexampleData [80], available at, https://bioconductor.org/packages/release/data/experiment/html/STexampleData.html. The full mouse hypothalamus data by MERFISH is available at, https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248, from which we focused on only "Replicate 6", as it had the largest number of cells and highest confluency [34]. The SCCOHT dataset by 10X Visium was collected at the University of Colorado Denver - Anschutz Medical Campus, and is available upon request.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Public Health Sciences, School of Medicine, Medical University of South Carolina, Charleston, USA. [2]Department of Obstetrics and Gynecology, School of Medicine, University of Colorado Denver - Anschutz Medical Campus, Aurora, USA. [3]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver - Anschutz Medical Campus, Aurora, USA.

References

1. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., *et al.*: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)
2. Shah, S., Lubeck, E., Zhou, W., Cai, L.: In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus. Neuron **92**(2), 342–357 (2016)
3. Shah, S., Lubeck, E., Zhou, W., Cai, L.: seqfish accurately detects transcripts in single cells and reveals robust spatial organization in the hippocampus. Neuron **94**(4), 752–758 (2017)
4. Wang, G., Moffitt, J.R., Zhuang, X.: Multiplexed imaging of high-density libraries of rnas with merfish and expansion microscopy. Scientific reports **8**(1), 1–13 (2018)
5. Xia, C., Fan, J., Emanuel, G., Hao, J., Zhuang, X.: Spatial transcriptome profiling by merfish reveals subcellular rna compartmentalization and cell cycle-dependent gene expression. Proceedings of the National Academy of Sciences **116**(39), 19490–19499 (2019)
6. Eng, C.-H.L., Lawson, M., Zhu, Q., Dries, R., Koulena, N., Takei, Y., Yun, J., Cronin, C., Karp, C., Yuan, G.-C., *et al.*: Transcriptome-scale super-resolved imaging in tissues by rna seqfish+. Nature **568**(7751), 235–239 (2019)
7. Asp, M., Bergenstråhle, J., Lundeberg, J.: Spatially resolved transcriptomes—next generation tools for tissue exploration. BioEssays **42**(10), 1900221 (2020)
8. Guilliams, M., Bonnardel, J., Haest, B., Vanderborght, B., Wagner, C., Remmerie, A., Bujko, A., Martens, L., Thoné, T., Browaeys, R., *et al.*: Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. Cell **185**(2), 379–396 (2022)
9. Moses, L., Pachter, L.: Museum of spatial transcriptomics. Nature Methods **19**(5), 534–546 (2022)
10. Dhainaut, M., Rose, S.A., Akturk, G., Wroblewska, A., Nielsen, S.R., Park, E.S., Buckup, M., Roudko, V., Pia, L., Sweeney, R., *et al.*: Spatial crispr genomics identifies regulators of the tumor microenvironment. Cell **185**(7), 1223–1239 (2022)
11. Chen, W.-T., Lu, A., Craessaerts, K., Pavie, B., Frigerio, C.S., Corthout, N., Qian, X., Laláková, J., Kühnemund, M., Voytyuk, I., *et al.*: Spatial transcriptomics and in situ sequencing to study alzheimer's disease. Cell **182**(4), 976–991 (2020)
12. Maynard, K.R., Collado-Torres, L., Weber, L.M., Uytingco, C., Barry, B.K., Williams, S.R., Catallini, J.L., Tran, M.N., Besich, Z., Tippani, M., *et al.*: Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. Nature neuroscience **24**(3), 425–436 (2021)
13. Ortiz, C., Carlén, M., Meletis, K.: Spatial transcriptomics: molecular maps of the mammalian brain. Annual review of neuroscience **44**, 547–562 (2021)
14. Levy-Jurgenson, A., Tekpli, X., Kristensen, V.N., Yakhini, Z.: Spatial transcriptomics inferred from pathology whole-slide images links tumor heterogeneity to survival in breast and lung cancer. Scientific reports **10**(1), 1–11 (2020)
15. Yoosuf, N., Navarro, J.F., Salmén, F., Ståhl, P.L., Daub, C.O.: Identification and transfer of spatial transcriptomics signatures for cancer diagnosis. Breast Cancer Research **22**(1), 1–10 (2020)
16. Hunter, M.V., Moncada, R., Weiss, J.M., Yanai, I., White, R.M.: Spatially resolved transcriptomics reveals the architecture of the tumor-microenvironment interface. Nature communications **12**(1), 1–16 (2021)
17. Kwon, S.: Single-molecule fluorescence in situ hybridization: quantitative imaging of single rna molecules. BMB reports **46**(2), 65 (2013)
18. Lubeck, E., Coskun, A.F., Zhiyentayev, T., Ahmad, M., Cai, L.: Single-cell in situ rna profiling by sequential hybridization. Nature methods **11**(4), 360–361 (2014)
19. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., Zhuang, X.: Spatially resolved, highly multiplexed rna profiling in single cells. Science **348**(6233), 6090 (2015)
20. Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F., Macosko, E.Z.: Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. Science **363**(6434), 1463–1467 (2019)

21. 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. Nature biotechnology **39**(3), 313–319 (2021)
22. Bergenstråhle, J., Larsson, L., Lundeberg, J.: Seamless integration of image and molecular analysis for spatial transcriptomics workflows. BMC genomics **21**(1), 1–7 (2020)
23. Vickovic, S., Eraslan, G., Salmén, F., Klughammer, J., Stenbeck, L., Schapiro, D., Äijö, T., Bonneau, R., Bergenstråhle, L., Navarro, J.F., *et al.*: High-definition spatial transcriptomics for in situ tissue profiling. Nature methods **16**(10), 987–990 (2019)
24. Atta, L., Fan, J.: Computational challenges and opportunities in spatially resolved transcriptomic data analysis. Nature Communications **12**(1), 1–5 (2021)
25. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J., Lundeberg, J.: Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage iii cutaneous malignant melanoma. Cancer research **78**(20), 5970–5979 (2018)
26. Navarro, J.F., Croteau, D.L., Jurek, A., Andrusivova, Z., Yang, B., Wang, Y., Ogedegbe, B., Riaz, T., Støen, M., Desler, C., *et al.*: Spatial transcriptomics reveals genes associated with dysregulated mitochondrial functions and stress signaling in alzheimer disease. Iscience **23**(10), 101556 (2020)
27. Kats, I., Vento-Tormo, R., Stegle, O.: Spatialde2: Fast and localized variance component analysis of spatial transcriptomics. bioRxiv (2021)
28. Rao, A., Barkley, D., França, G.S., Yanai, I.: Exploring tissue architecture using spatial transcriptomics. Nature **596**(7871), 211–220 (2021)
29. Jiang, X., Xiao, G., Li, Q.: A bayesian modified ising model for identifying spatially variable genes from spatial transcriptomics data. Statistics in Medicine **41**(23), 4647–4665 (2022)
30. Wang, Y., Ma, S., Ruzzo, W.L.: Spatial modeling of prostate cancer metabolic gene expression reveals extensive heterogeneity and selective vulnerabilities. Scientific reports **10**(1), 1–14 (2020)
31. Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidi, A., Vickovic, S., Larsson, L., *et al.*: Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. Nature communications **9**(1), 2419 (2018)
32. Cable, D.M., Murray, E., Zou, L.S., Goeva, A., Macosko, E.Z., Chen, F., Irizarry, R.A.: Robust decomposition of cell type mixtures in spatial transcriptomics. Nature Biotechnology **40**(4), 517–526 (2022)
33. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M., *et al.*: Integrated analysis of multimodal single-cell data. Cell **184**(13), 3573–3587 (2021)
34. Svensson, V., Teichmann, S.A., Stegle, O.: Spatialde: identification of spatially variable genes. Nature methods **15**(5), 343–346 (2018)
35. 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. Molecular Therapy-Nucleic Acids (2021)
36. Edsgärd, D., Johnsson, P., Sandberg, R.: Identification of spatial expression trends in single-cell gene expression data. Nature methods **15**(5), 339–342 (2018)
37. Sun, S., Zhu, J., Zhou, X.: Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. Nature methods **17**(2), 193–200 (2020)
38. Zhu, J., Sun, S., Zhou, X.: Spark-x: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. Genome Biology **22**(1), 1–25 (2021)
39. 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. Nature biotechnology **36**(12), 1183–1190 (2018)
40. Miller, B.F., Bambah-Mukku, D., Dulac, C., Zhuang, X., Fan, J.: Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. Genome research **31**(10), 1843–1855 (2021)
41. 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. Nature methods **18**(11), 1342–1351 (2021)
42. Liu, Q., Hsu, C.-Y., Shyr, Y.: Scalable and model-free detection of spatial patterns and colocalization. Genome research **32**(9), 1736–1745 (2022)
43. Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H.: Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(4), 825–848 (2008)
44. Liu, D., Lin, X., Ghosh, D.: Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics **63**(4), 1079–1088 (2007)
45. Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., Heckerman, D.: Fast linear mixed models for genome-wide association studies. Nature methods **8**(10), 833–835 (2011)
46. Christensen, O.F., Waagepetersen, R.: Bayesian prediction of spatial count data using generalized linear mixed models. Biometrics **58**(2), 280–286 (2002)
47. Dean, C.B., Ugarte, M.D., Militino, A.F.: Penalized quasi-likelihood with spatially correlated data. Computational statistics & data analysis **45**(2), 235–248 (2004)
48. Zhang, K., Peters, J., Janzing, D., Schölkopf, B.: Kernel-based conditional independence test and application in causal discovery. arXiv preprint arXiv:1202.3775 (2012)
49. Rencher, A.C.: A review of "Methods of Multivariate Analysis, ". Taylor & Francis (2005)
50. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International journal of computer vision **40**(2), 99–121 (2000)
51. Odén, A., Wedel, H.: Arguments for fisher's permutation test. The Annals of Statistics, 518–520 (1975)
52. Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A., Dulac, C., *et al.*: Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. Science **362**(6416), 5324 (2018)

53. Kirsch, L., Liscovitch, N., Chechik, G.: Localizing genes to cerebellar layers by classifying ish images. PLOS computational biology **8**(12), 1002790 (2012)
54. Sanders, B.E., Wolsky, R., Doughty, E.S., Wells, K.L., Ghosh, D., Ku, L., Pressey, J.G., Bitler, B.B., Brubaker, L.W.: Small cell carcinoma of the ovary hypercalcemic type (sccoht): A review and novel case with dual germline smarca4 and brca2 mutations. Gynecologic Oncology Reports, 101077 (2022)
55. Zhang, L., Wang, Y., Sha, Y., Zhang, B., Zhang, R., Zhang, H., Xu, S., Wang, H., Xu, Y., Chen, Y., *et al.*: Cited4 enhances the metastatic potential of lung adenocarcinoma. Thoracic Cancer **12**(9), 1291–1302 (2021)
56. Gao, C., Guo, X., Xue, A., Ruan, Y., Wang, H., Gao, X.: High intratumoral expression of eif4a1 promotes epithelial-to-mesenchymal transition and predicts unfavorable prognosis in gastric cancer. Acta Biochimica et Biophysica Sinica **52**(3), 310–319 (2020)
57. Gan, L., Yang, Y., Li, Q., Feng, Y., Liu, T., Guo, W.: Epigenetic regulation of cancer progression by ezh2: from biological insights to therapeutic potential. Biomarker research **6**(1), 1–10 (2018)
58. Duan, R., Du, W., Guo, W.: Ezh2: a novel target for cancer treatment. Journal of hematology & oncology **13**(1), 1–12 (2020)
59. Chen, F., Qin, T., Zhang, Y., Wei, L., Dang, Y., Liu, P., Jin, W.: Reclassification of endometrial cancer and identification of key genes based on neural-related genes. Frontiers in Oncology **12** (2022)
60. Nelson, T.S., Taylor, B.K.: Targeting spinal neuropeptide y1 receptor-expressing interneurons to alleviate chronic pain and itch. Progress in neurobiology **196**, 101894 (2021)
61. Viswanathan, S., Bandyopadhyay, S., Kao, J.P., Kanold, P.O.: Changing microcircuits in the subplate of the developing cortex. Journal of Neuroscience **32**(5), 1589–1601 (2012)
62. Wang, Y.-X., Zhang, Y.-J.: Nonnegative matrix factorization: A comprehensive review. IEEE Transactions on knowledge and data engineering **25**(6), 1336–1353 (2012)
63. Seal, S., Wrobel, J., Johnson, A.M., Nemenoff, R.A., Schenk, E.L., Bitler, B.G., Jordan, K.R., Ghosh, D.: On clustering for cell phenotyping in multiplex immunohistochemistry (mihc) and multiplexed ion beam imaging (mibi) data. BMC Research Notes **15**(1), 215 (2022)
64. Seal, S., Ghosh, D.: MIAMI: mutual information-based analysis of multiplex imaging data. Bioinformatics **38**(15), 3818–3826 (2022). doi:10.1093/bioinformatics/btac414. https://academic.oup.com/bioinformatics/article-pdf/38/15/3818/45213227/btac414.pdf
65. Seal, S., Datta, A., Basu, S.: Efficient estimation of snp heritability using gaussian predictive process in large scale cohort studies. PLoS genetics **18**(4), 1010151 (2022)
66. CKI, R.C.W.: Gaussian processes for machine learning. International Journal of Neural Systems **14** (2006)
67. Gourieroux, C., Holly, A., Monfort, A.: Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. Econometrica: journal of the Econometric Society, 63–80 (1982)
68. Lea, A.J., Tung, J., Zhou, X.: A flexible, efficient binomial mixed model for identifying differential dna methylation in bisulfite sequencing data. PLoS genetics **11**(11), 1005650 (2015)
69. Sun, S., Hood, M., Scott, L., Peng, Q., Mukherjee, S., Tung, J., Zhou, X.: Differential expression analysis for rnaseq using poisson mixed models. Nucleic acids research **45**(11), 106–106 (2017)
70. Rao, C.R.: In: Balakrishnan, N., Nagaraja, H.N., Kannan, N. (eds.) Score Test: Historical Review and Recent Developments, pp. 3–20. Birkhäuser Boston, Boston, MA (2005)
71. Sun, S., Zhu, J., Mozaffari, S., Ober, C., Chen, M., Zhou, X.: Heritability estimation and differential analysis of count data with generalized linear mixed models in genomic sequencing studies. Bioinformatics **35**(3), 487–496 (2019)
72. Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., Smola, A.: A kernel statistical test of independence. Advances in neural information processing systems **20** (2007)
73. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances (2007)
74. Liu, Y., Chen, S., Li, Z., Morrison, A.C., Boerwinkle, E., Lin, X.: Acat: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. The American Journal of Human Genetics **104**(3), 410–421 (2019)
75. Cressie, N.: Statistics for spatial data. John Wiley & Sons (2015)
76. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165–1188 (2001)
77. Narum, S.R.: Beyond bonferroni: less conservative analyses for conservation genetics. Conservation genetics **7**, 783–787 (2006)
78. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**(43), 15545–15550 (2005)
79. Baddeley, A., Bárány, I., Schneider, R.: Spatial point processes and their applications. Stochastic Geometry: Lectures Given at the CIME Summer School Held in Martina Franca, Italy, September 13–18, 2004, 1–75 (2007)
80. Righelli, D., Weber, L.M., Crowell, H.L., Pardo, B., Collado-Torres, L., Ghazanfar, S., Lun, A.T., Hicks, S.C., Risso, D.: Spatialexperiment: infrastructure for spatially-resolved transcriptomics data in r using bioconductor. Bioinformatics **38**(11), 3128–3131 (2022)