SPRI: Spatial Pattern Recognition using Information based method for spatial gene expression data

Jing-Xian Hu [1], *Ye Yuan* [1,*], *and Hong-Bin Shen* [1,*]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

*Corresponding author

Address correspondence to Y.Y at: yuanye_auto@sjtu.edu.cn, H.B. Shen at: hbshen@sjtu.edu.cn

Abstract

The rapid development of spatially resolved transcriptomics has made it possible to analyze spatial gene expression patterns in complex biological tissues. To identify such genes, we propose a novel and robust nonparametric information-based approach, SPRI, to recognize their spatial patterns. SPRI directly models spatial transcriptome raw count data without model assumptions, which transforms the problem of spatial expression pattern recognition into the detection of dependencies between spatial coordinate pairs with gene read count as the observed frequencies. SPRI was used to analyze four recent published spatially resolved transcriptome data, and all results showed that SPRI outperforms prior methods, by robustly detecting more genes with significant spatial expression patterns, and revealing biological insights that cannot be identified by other methods.

Introduction

In recent years, the rapid development of high-throughput spatial transcriptome technologies enables the understanding of spatially resolved gene expression patterns in complicated tissues [1, 2]. Some of them are based on fluorescence in situ hybridization (FISH), which can locate each RNA transcripts in the sample [3, 4]. Others are based on sequencing technology, including spatial transcriptome (ST), Slide-Seq, 10x Visium, etc [5-7]. This technology first partitions tissue into small regions (spots) to associates all transcripts within one spot with known spatial coordinate barcodes, and then sequences them to capture the expression levels of thousands of genes in the spot. Such technology provides an efficient spatial approach for new biological discoveries and understanding of multiple biological processes in disease [8, 9].

Identification of genes with spatial expression patterns (SE genes) is an essential step in analysis of spatial transcriptome data. For this task, the several existing methods can be divided into two groups: normalized data based method and raw count data based method. Trendsceek [10] uses a two-dimensional point process to describe the spatial location distribution of cells, while gene expression levels are described by a probability

distribution of scalar values. SpatialDE [11] constructs multidimensional Gaussian distribution for normalized gene expression. MERINGUE [12] is based on spatial autocorrelation and cross-correlation for normalized gene expression. BinSpect [13] is based on enrichment analysis of spatial network neighbors in binarized high gene expression state. Without normalization, SPARK [14] uses a generalized linear spatial model with a series of custom spatial kernel functions to describe the raw count data using Poisson distribution. However, these existing methods still have limitations: 1) most of them are based on normalized gene expression data, thus fail to consider the variance in raw counts. 2) Prior methods are based on certain statistical assumptions that limit their ability to identify various possible spatial distribution patterns. For example, Trendsceek focuses on modeling two points in space; SpatialDE assumes that the data obey Gaussian distribution; MERINGUE and BinSpect focus on modeling spatial neighbors, assuming that differences between neighbors are comparable; and SPARK requires settings of specific spatial kernel functions. 3) They only assign significance to rank genes, however low P or Q values do not necessarily mean real spatial patterns [15].

In this work, we propose nonparametric **S**patial **P**attern **R**ecognition using **I**nformation based method, SPRI, which models raw count data directly without model assumptions to give the rank of gene spatial expression patterns. SPRI firstly converts the spatial gene pattern problem into an association detection problem between (x, y) coordinate values with observed raw count data, and then estimates associations using an information-based method, TIC [16, 17], which calculates the total mutual information with all possible x-y-grids. Without unnecessary assumptions, SPRI can detect more SE gene patterns with higher accuracy.

Results

Simulations.

The overview of SPRI is shown in **Fig. 1a**. Unlike Trendsceek, SpatialDE, MERINGUE and BinSpect, which are based on normalized gene expression data with assumption that the sum of RNA transcripts of each cell is equal, SPRI directly models

the raw count data. Unlike SPARK, which is based on statistical hypothesis of Gaussian distribution and certain spatial gene pattern kernel assumptions, SPRI converts the spatial gene pattern problem to association detection problem between coordinates values of (x, y) using observed count data as observed frequencies, and it then estimates the association using the information-based approach, TIC to calculates the total mutual information with all possible x-y-grids. Without these unnecessary assumptions, SPRI can detect more SE gene patterns theoretically. To evaluate the performance of SPRI, we compared it with five recently developed methods with precision plots, including SPARK, SpatialDE, Trendsceek, MERINGUE and BinSpect on four simulated data (**Fig. 1b, Supplementary Fig. 1**).

Following comparison strategy in ref. [10, 14], the simulated patterns are set as Hotspot, Streak, Step gradient and Linear gradient respectively. To explore the robustness of these methods, we also tried different parameters for the simulation data to estimate the standard deviation in the plot. See the details in **Supplementary Notes**. As can be seen, for all four simulation patterns, SPRI outperforms all other prior methods on the task of identifying spatial expression (SE) genes. Among these methods, SPARK, MERINGUE and BinSpect is the second best one followed by SpatialDE and Trendsceek on different simulated patterns respectively, which is consistent with previous studies [12-14].

Mouse olfactory bulb data (MOB Replicate 11).

The first dataset we used to test SPRI is replicate layer 11 of mouse olfactory bulb (MOB Replicate 11) [7], which has 16,218 genes measured on 260 spots. SPRI ranks the genes using TIC scores. To test the significance of the top-ranked genes, we performed a permutation test to compute P value for the top 10% genes and then used FDR correction to compare with existing methods, including SPARK, SpatialDE, Trendsceek, MERINGUE and BinSpect. As can be seen in **Fig. 2a**, SPRI can identify more potential genes. Following SPARK paper, we named the genes with a FDR cutoff of 0.05 as SE genes. For MOB Replicate 11 data, SPRI identified 1,102 genes, while SPARK identified 772 genes (overlap with SPRI= 312; **Supplementary Fig. 2a**),

SpatialDE identified 67 genes (overlap with SPRI = 47), Trendsceek did not identified any SE genes, MERINGUE identified 720 genes (overlap with SPRI= 315), and BinSpect identified 804 genes (overlap with SPRI= 309).

We firstly compared the SE genes identified by SPRI with known marker gene list to further validate our method. A list of 2,030 cell type-specific marker genes was downloaded from a recent single-cell RNA sequencing research of olfactory bulbs [18]. Fisher's exact test was used to quantify the gene overlap. As shown in **Fig. 2b**, SPRI demonstrates higher enrichment than other methods. Secondly, the proportion of top SPARK-ranked SE genes that were also identified by SPRI as SE genes and the proportion of top SPRI-ranked genes that were also identified by SPARK as SE genes were compared (**Fig. 2c**). The results showed that SPRI can covers more top SPARK ranked genes. The comparison with other methods can be found in **Supplementary Fig. 2d**. Thirdly, the comparison of expression levels for SE genes shows that SPRI SE genes can detect more highly expressed genes than existing methods. As shown in **Fig. 2d**, the expression level uniquely detected by SpatialDE was close to zero, and the level uniquely detected by SPARK, MERINGUE and BinSpect are comparable. In contrast, the level of SE genes identified by SPRI only is closest to that of SE genes found by all five methods, which is the highest.

To visually evaluate the SE genes detected by SPRI, we also clustered the 1,102 SE genes identified by SPRI and obtained five major spatial patterns (**Supplementary Fig. 2f**). The first three patterns correspond to three cell layers of mouse olfactory layer respectively: mitral cell layer (pattern I), glomerular layer (pattern II), and the granular cell layer (pattern III). The top SPRI-ranked genes were selected to visualize these three spatial patterns (**Fig. 2e**), of which *Scg2* [19] and *Gabrb3* [20] were identified only by SPRI. The *in situ* hybridization images from the Allen Brain Atlas further cross-validated these genes exhibiting spatial expression patterns (**Fig. 2f**).

We next explore the biological insights found by SPRI. Manual inspection of the top five SE genes uniquely identified by SPRI and other methods (**Supplementary Fig. 3)** indicates that SPRI genes are more spatially variable, and all of them are found associated with brain functions, supported by literature, including *Cst3*, *Fth1*, *Mdh1*,

Rtn4 and *Ddx5*. For example, the B/B polymorphism of *Cst3* can lead to reduced secretion of cystatin C and decreased efficiency of signal peptide cleavage, which in turn increases the risk of Alzheimer's disease [21]. *Fth1* was found to be associated with ferritinophagy and ferroptosis, which is an important regulatory mechanism in Parkinson's disease [22]. *Mdh1* is a key bioenergetic protein in the TCA cycle of the mouse brain, which is irreversibly oxidized and accumulated in the aged brain [23]. *Rtn4* is a mylein-associated glycoprotein, and studies have shown that knockdown of *Rtn4* would causes symptoms of schizophrenia-like behavior [24]. *Ddx5* acts as a transcriptional regulator of LINC01116 to the IL-1β promoter, activating IL-1β expression to promote glioma proliferation [25]. In addition, functional enrichment analyses of SE genes detected by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect was also performed (**Methods**). We firstly compared the top 10 Gene Ontology (GO) terms found by these five methods for the same number of genes (top 100 (**Fig. 2g**), 150 and 200 in **Supplementary Fig. 2g**), which indicates that SPRI obtains much more significant GO terms than other methods. Then, functional enrichment analyses were performed on whole SE genes at 0.05 FDR cutoff (**Fig. 2h**). Totally, 1,280 GO terms and 84 Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were enriched in the SE genes identified by SPRI, while SPARK had 1,157 enriched GO terms (overlap with SPRI = 663; **Supplementary Fig. 2b**) and 83 KEGG pathways (overlap = 46; **Supplementary Fig. 2c**), SpatialDE had only 99 (overlap = 60) enriched GO terms and 2 KEGG pathways (overlap = 2), MERINGUE had 1,112 (overlap= 630) enriched GO terms and 83 KEGG pathways (overlap = 50), and BinSpect had 788 (overlap = 531) enriched GO terms and 38 KEGG pathways (overlap = 32). The result shows that many enriched GO terms detected by SPRI only are associated with synaptogenesis and olfactory bulb development, such as synaptic vesicle localization (GO 0097479; SPRI *P* value= 2.5×10^{-6}, SPARK *P* value = 9.39×10^{-3}, MERINGUE *P* value = 1.68×10^{-3}, BinSpect *P* value = 3.71×10^{-3}, while SpatialDE did not has this enriched GO term). In addition, many KEGG pathways identified by SPRI only are directly relevant to nervous system disease, such as Parkinson disease (KEGG mmu05012; SPRI *P* value = 1.21×10^{-36}, SPARK *P* value = 1.68×10^{-1}, MERINGUE *P*

value = 6.93×10^{-2}, BinSpect *P* value = 5.36×10^{-2}). An additional functional enrichment analysis was also performed on SE genes identified by SPRI only, of which the result is consistent with that of all SPRI SE genes (**Supplementary Fig. 2h**).

Mouse olfactory bulb data (MOB Replicate 12)

The second dataset is replicate layer 12 of mouse olfactory bulb (MOB Replicate 12) [7], which has 16,034 genes measured on 282 spots. As can be seen in **Fig. 3a**, SPRI identified more potential genes within a certain range of FDRs. For MOB Replicate 12 data, SPRI identified 1,565 genes, while SPARK identified 519 genes (overlap with SPRI = 302; **Supplementary Fig. 4a**), SpatialDE identified 285 genes (overlap with SPRI = 184), Trendsceek only identified 46 SE genes, MERINGUE identified 523 genes (overlap with SPRI = 317), BinSpect identified 573 genes (overlap with SPRI = 274).

Similar to the result of **Fig. 2**, the comparison of expression levels for SE genes shows that SPRI SE genes can detect more highly expressed genes than existing methods (**Fig. 3b**). Secondly, **Fig. 3c** shows that SPRI can cover most top SPARK SE genes. More results can be found in **Supplementary Fig.4d**. Thirdly, SE genes identified by SPRI only were highly enriched in the same marker gene list [18] (**Fig. 3d**).

We also clustered the SE genes identified by SPRI and obtained five major spatial patterns (**Supplementary Fig.4f**). The same three patterns with corresponding glomerular layer (pattern V), and the granular cell layer (pattern II), mitral cell layer (pattern III), were visualized by top SPRI-ranked genes (**Fig. 3e**), which were also cross-validated by the *in situ* hybridization images from the Allen Brain Atlas (**Fig. 3f**).

We next evaluate the biological insights found by SPRI. Manual inspection of the top five SE genes uniquely identified by SPRI (**Supplementary Fig. 5)** indicates that all of them are found associated with mouse olfactory bulb development. In addition to *Cst3* and *Ddx5* we have discussed, *Actb*, *Tuba1a* and *Rplp1* were also associated with brain activity. Actin beta (*Actb*), a structural backbone housekeeping protein, supports accelerated axonal growth when its putative functionally acquired missense mutation leads to human Baraitser-Winter syndrome, characterized by mental retardation,

cortical malformations, and sensorineural deafness [26]. Mutations in *Tuba1a*, the major alpha-tubulin expressed during brain development, cause a range of human brain malformation disorders [27]. *Rplp1* was found to be a ribosomal protein essential for brain development and cell proliferation [28]. Finally, functional enrichment analyses were performed. The top 10 GO terms found by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect for top genes (100 (**Fig. 3g**), 150 and 200; **Supplementary Fig. 4g**) were compared.

Since MOB Replicate 11 and MOB Replicate 12 are two different layers of the same experiment, to evaluate their robustness, we calculated the overlap rate for the top 10 GO terms enriched by the top SE genes on both data (**Fig. 3h**). Functional enrichment analyses were also performed on all SE genes at 0.05 cutoff of FDR, for SPRI (**Fig. 3i**), SPARK, SpatialDE, MERINGUE and BinSpect (**Supplementary Fig. 4b** and **c**). A lot of enriched GO terms detected by SPRI only are associated with synaptogenesis and olfactory bulb development, such as structural constituent of synapse (GO 0097479; SPRI *P* value = 2.47×10^{-7}, SPARK *P* value = 1.70×10^{-2}, SpatialDE *P* = 2.61×10^{-2}, MERINGUE *P* = 4.58×10^{-3}, BinSpect did not has this enriched GO term). In addition, many KEGG pathways identified by SPRI only are directly relevant to nervous system disease, such as Huntington disease (SPRI *P* value = 6.03×10^{-37}, MERINGUE *P* value = 5.84×10^{-2}, BinSpect *P* value = 5.81×10^{-2}, while SPARK and SpatialDE did not has this enriched KEGG pathways). An enrichment analysis was also performed on SE genes identified by SPRI only (**Supplementary Fig. 4h**).

Human breast cancer data (Breast Cancer Layer 2)

The third dataset is layer 2 of breast cancer (breast cancer layer 2) [7], which has 14,789 genes measured on 250 spots. SPRI identified more potential genes within a certain range of FDRs (**Fig. 4a)**. Totally, SPRI identified 1,151 SE genes, while SPARK identified only 290 SE genes (overlap with SPRI = 212; **Supplementary Fig. 6a**), SpatialDE identified 115 SE genes (overlap with SPRI = 59), Trendsceek only identified 13 SE genes, MERINGUE identified 207 SE genes (overlap with SPRI =184), BinSpect identified 146 SE genes (overlap with SPRI = 100).

Still, the expression comparison of SE genes shows that SPRI can detect more highly expressed genes than existing methods (**Fig. 4b**). Secondly, SPRI can cover most top SPARK SE genes (**Fig. 4c**), more results can be found in **Supplementary Fig. 6d**. We also compared the SE genes identified by SPRI with known marker gene list to further validate our method (**Fig. 4d**). The list of genes related to human breast cancer was downloaded from CancerMine database [29]. Top SPRI-ranked genes only identified by SPRI were also listed (**Fig. 4e**) to visually evaluate the correctness of SE genes detected by SPRI.

We next evaluate the biological insights found by SPRI. Manual inspection of the top five SE genes uniquely identified by SPRI (**Supplementary Fig. 7)** indicates that four of them are found associated with breast cancer, including *ACTB*, *TMSB10*, *PABPC1* and *ACTG1*. Study finds differential *ACTB* expression in breast cancer is associated with metastasis and drug resistance in breast cancer [30]. *TMSB10* was upregulated in breast cancer tissues and its overexpression promotes invasion, proliferation and migration of breast cancer cells [31]. The *PABPC1* gene was a downstream target of *SNHG14* and mediates *SNHG14-induced* oncogenesis in breast cancer [32]. *ACTG1*, a cytoskeletal protein, is thought to be a component of the cell migration machinery, and when destabilized is able to inhibit the migration of cancer cells [33]. In addition, functional enrichment analyses were performed. We firstly compared the top 10 GO terms found by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect for the same number of genes (top 100 (**Fig. 4f**), 150 and 200; **Supplementary Fig. 6e**). Then, functional enrichment analyses were performed on all SE genes at 0.05 FDR cutoff for SPRI (**Fig. 4g**), SPARK, SpatialDE, MERINGUE and BinSpect (**Supplementary Fig. 6b** and **c**), which shows that many enriched GO terms detected by SPRI only are associated with immune responses, such as synaptic vesicle localization (GO 0002433; SPRI *P* value= 5.44×10^{-5}, while SPARK, SpatialDE, MERINGUE and BinSpect did not has this enriched GO term).

Human breast cancer data (BC23209_C1_stdata)

The last dataset is BC23209_C1_stdata of breast cancer (breast cancer layer 2) [5],

which has 16,859 genes measured on 294 spots. SPRI identified 812 genes, while SPARK identified 142 genes, SpatialDE identified 210 genes, Trendsceek identified 216 SE genes, MERINGUE identified 215 SE genes and BinSpect identified 9 SE genes, respectively **(Fig. 5a, Supplementary Fig. 8a**). Consistent with previous analysis, SPRI can detect more highly expressed SE genes (**Fig. 5b**) and can cover most top ranked SE genes (**Fig. 5c, Supplementary Fig. 8d**). We also compared the SE genes identified by SPRI with the same marker gene list [29] related human breast cancer (**Fig. 5d**).

After visualization of SE genes detected by SPRI (**Fig. 5e**), we evaluate the biological insights found by SPRI. Manual inspection of the top five SE genes uniquely identified by SPRI (**Supplementary Fig. 9)** indicates that three of them have been found associated with breast cancer, including, *RPS21*, *PPP1CA* and *TXNIP*. *RPS21*'s role in breast cancer is not clear now, but one transcript of *RPS*, *AA-RPS21*, is differentially expressed in cancerous tissues, indicating its potential driver role in breast cancer [34]. *PPP1CA* together with *PRKACG* and *PRKAR1B* were found to be the most strongly associated with breast cancer–specific survival [35]. Study found that inhibition of *TXNIP* via *Myc* drove Triple-negative breast cancers aggressiveness and was associated with decreased metastasis-free survival and decreased overall survival in breast cancer [36]. For functional enrichment analyses, we firstly compared the top 10 GO terms found by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect for the same number of genes (top 100 (**Fig. 5f**), 150 and 200; **Supplementary Fig. 8e**). Then, functional enrichment analyses were performed on all SE genes at 0.05 FDR cutoff for SPRI (**Fig. 5g**), SPARK, SpatialDE, MERINGUE and BinSpect (**Supplementary Fig. 8b** and **c**). Many enriched GO terms detected by SPRI only are associated with immune responses, such as posttranslational protein targeting to endoplasmic reticulum membrane (GO 0006620; SPRI *P* value= 5.01×10^{-4}, while SPARK, SpatialDE, MERINGUE and BinSpect did not has this enriched GO term). An additional functional enrichment analysis was also performed on SE genes identified by SPRI only (**Supplementary Fig. 8f**).

Discussion

The recent rapid development of high throughput spatial transcriptomics technology opens a door how to understand the spatial resolved biological behaviors of genes and cells. One essential and initial step of such analysis is to detect genes with spatial expression patterns.

In this work, we propose a novel information-based spatial pattern gene identification method, SPRI, to model spatial raw count data directly. It converts the SE gene detection problem to a dependencies mining problem between spatial coordinate pairs with raw gene read count as the observed frequencies. Such strategy distinguishes SPRI from prior existing SE methods relying on certain model or assumptions. For example, methods based on normalization data assume implicitly that the total number of RNA transcripts is identical, which is not always true [37]. Other methods modeling raw count, like SPARK, also rely on certain parametric statistical model/hypothesis or designed kernel functions, which still limits the ability to identify various possible spatial distribution patterns.

To evaluate SPRI's performance, we compared it with five existing methods on four publicly available datasets comprehensively. The results consistently indicate that SPRI can robustly identify more genes with true spatial expression patterns validated by *In situ* hybridization experiments, and that SE genes identified by SPRI uniquely are more spatially variable and are supported by recent studies.

Methods

SPRI: model and algorithm

In this work, we convert the problem of identifying genes with spatial expression patterns into the problem of identifying dependencies on (X, Y) coordinate observations based on the raw count expression of genes in the two-dimensional space of cells/spots.

(a) **Computing the total information coefficient for each gene**. The idea of TIC [17] is based on MIC [16], in which the range of two variables is partitioned by a grid to

evaluate if there is a dependency between the two variables. Specifically, given a set of two jointly observed data (x, y) for variables (X, Y), the mutual information $I\big((X,Y),k,l\big)$ is computed under all k-by-l grids:

$$I\big((X,Y),k,l\big) = \sum_{y\in Y}\sum_{x\in X} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) \tag{1}$$

where the number of grids $kl < B$. B depends on the number of samples and is usually set by $B = n^{0.6}$. For each k-by-l grid, the maximum mutual information is retained:

$$I^*\big((X,Y),k,l\big) = \max_{G\in G(k,l)} I((X,Y)|_G) \tag{2}$$

For a fair comparison under different grid divisions, the maximum mutual information $m_{k,l}$ under each grid G are normalized to between 0 and 1, constituting characteristic matrix $M = (m_{k,l})$.

$$\widehat{M}(D)_{k,l} = \frac{I^*(D,k,l)}{\log min\{k,l\}} \tag{3}$$

MIC and TIC are two different properties of the characteristic matrix. MIC is the supremum value in M, while TIC is the sum of M. Compared to MIC, which only considers only the maximal value of the characteristic matrix and may throw away meaningful information, TIC is able to obtain a smaller bias and better power by summing over all entries in the independent case. In other words, TIC is able to measure the presence or absence of dependencies between two variables. The definitions of MIC and TIC are as follow:

$$MIC(D) = \max \widehat{M}(D)_{k,l} \tag{4}$$

$$TIC(D) = \sum_{kl\leq B(n)} \widehat{M}(D)_{k,l} \tag{5}$$

We apply TIC to determine whether there is a dependency relationship between two-variable (X, Y) to identify genes with spatial expression patterns, and the higher the TIC

value, the more likely the gene has a spatial expression pattern. For all genes we gave a SE gene ranking based on TIC values.

(b) **Background correction**. Considering the effect of the spatial shape of the tissue, we added uniformly distributed background spot locations (**Fig. 1a**) to convert the shape to rectangle while preserving the original spot locations. The expression of the background is set as the statistical mean value [38] of the gene as follow:

$$mean = \frac{1}{2}Q_2 + \frac{1}{4}(Q_1 + Q_3) \tag{6}$$

where Q_1, Q_2, and Q_3 are the first, second and third quartiles of current gene expression respectively.

(c) **Identifying statistically significant SE genes with permutation test.** To compare with other algorithms, we performed a permutation test on the top 10% of genes in the TIC ranking. In the permutation test, we keep the spot locations fixed, randomly disrupt gene expression, and then recalculate the TIC values. The *P*value of each gene is computed as follow:

$$P\ value = \frac{\{m|TIC_i > TIC_{true}, i=1,2,3\ldots,M\}}{N} \tag{7}$$

where TIC_i is the TIC value for the *i-th* permutation, and TIC_{true} is the original TIC value of gene expression; *N* is the number of total permutation times, we set *N* to 10,000 in our experiment. After getting the *P* values for each gene, we used Benjamini-Yekutieli in the python package of "statsmodels" to control FDR. The genes with FDR <0.05 are considered as significant SE genes.

Gene sets, cluster, visualization and functional enrichment analysis

For these spatial transcriptomics datasets, we downloaded lists of known genes to validate the SE genes recognized by different methods. For the mouse olfactory bulb

data, we downloaded a list from of 2,030 cell type-specific marker genes identified in recent single-cell RNA sequencing studies in the olfactory bulb [18]. For the human breast cancer data, we downloaded a list of genes associated with breast cancer from the CancerMine database (http://bionlp.bcgsc.ca/cancermine/). These breast cancer related genes are composed of three parts, namely cancer drivers, oncogenes, and tumor suppressors. SE genes were validated by these gene lists respectively. Besides, the clustering code provided by SPARK author was directly used to cluster SE genes (FDR<0.05) identified by SPRI to five clusters following SPARK paper. Furthermore, we follow the visualization strategy of gene spatial expression patterns proposed by SPARK and the raw count was directly used in Supplementary Materials. In SPARK, variance-stabilizing transformation was performed on the raw count data and the log-scale total counts was adjusted to get a relative gene-expression. Finally, the same number of top SE genes and whole SE genes at a 0.05 FDR cutoff identified by SPRI and SPARK were used for functional enrichment analysis including GO terms analysis and KEGG pathways analysis. Following the SPARK paper, we adopted the R package of "clusterProfiler (v3.18.1)" to perform all functional enrichment analysis. In the package, we set the *P* value correction method as the default 'BH' and the cutoff of FDR as 0.05.

Spatial transcriptomics datasets.

In this work, four spatial transcriptomics datasets including mouse olfactory bulb data and two human breast cancer data, were downloaded from Spatial Research (https://www.spatialresearch.org/). These spatial transcriptomics datasets consist of two components: the spatial locations(spots) and the gene expression (read counts) observed at these spatial locations.

For the two mouse olfactory bulb datasets, we adopted the 'MOB Replicate 11' file and 'MOB Replicate 12' file of [7] for analysis. 'MOB Replicate 11' file contains 16,218 genes observed on 262 spots and 'MOB Replicate 12' file contains 16,034 genes observed on 282 spots. For the two human breast cancer datasets, we adopted the

‘Breast Cancer Layer 2’ file of [7] and ‘BC23209_C1_stdata’ file of [5] for analysis. ‘Breast Cancer Layer 2’ file contains 14,789 genes observed on 251 spots and ‘BC23209_C1_stdata’ file contains 16,859 genes observed on 294 spots. Following the SPARK paper, we filtered out spots less than ten total read counts. Through data filtering, we finally analyzed on 260 spots in ‘MOB Replicate 11’ data, 279 spots in ‘MOB Replicate 12’ data, 250 spots in ‘Breast Cancer Layer 2’ data, 294 spots in ‘BC23209_C1_stdata’ data. Then we select the top 10% of TIC ranked genes for permutation test.

Comparison of methods

We compare our method SPRI with five prior algorithms for spatial expression pattern recognition of genes, including SPARK [14], SpatialDE [11], Trendsceek [10], MERINGUE [12] and BinSpect [13]. SpatialDE, Trendsceek, MERINGUE and BinSpect are based on normalized data and SPARK is based on raw count data.

The first method we compared with is the SPARK (R package SPARK; v1.1.0), we directly use SPARK's code on github for analysis (https://github.com/xzhoulab/SPARK-Analysis). Following the SPARK paper, we performed the same data preprocessing on the four spatial transcriptomic data. Specifically, genes that are expressed in less than 10% of the spots were filtered out, and only spots containing at least ten total read counts were retained. According to the SPARK paper, if the adjusted *P* value (i.e., FDR) output by SPARK is below the threshold of 0.05, the identified SE is significant.

The second method we compared with is the SpatialDE (python package SpatialDE; v.1.1.3), we directly use SpatialDE’s code downloaded from github for analysis (https://github.com/Teichlab/SpatialDE). Following the SpatialDE paper, the SE gene is considered significant if the *Q* value (i.e.,FDR) was below the threshold of 0.05.

The third method we compared with is the Trendsceek (R package trendsceek; v.1.0.0). We directly used the code provided from github for analysis (https://github.com/edsgard/trendsceek). Following the Trendsceek paper, we

performed the same data preprocessing on the four spatial transcriptomic data. Specifically, genes that express less than three spots were filtered out, and only spots containing at least five read counts were retained. Then the raw count data were processed through log10 transformation. For the real data, the top 500 variable genes were taken for analysis. The SE gene is considered significant if the p.bh value (i.e.,FDR) was below the threshold of 0.05.

The fourth method we compared with is MERINGUE (R package MERINGUE; v.1.0). The code provided on github was used for analysis (https://github.com/JEFworks-Lab/MERINGUE). Following MERINGUE paper, poor spots (fewer than 100 read counts) and poor genes (fewer than 100 read counts) were filtered out. For BC23209_C1_stdata data, poor spots (fewer than 1 read counts) and poor genes (fewer than 1 read counts) were filtered out. Then Benjamini–Yekutieli correction was performed to control FDR, the SE gene is regarded significant if the FDR was below the threshold of 0.05.

The last method we compared with is the BinSpect (R package Giotto; v.1.1.0). The code provided from github was used for analysis (https://rubd.github.io/Giotto_site/). BinSpect provide two different ways in binarization, BinSpect-kmens and BinSpect-rank. Since the difference between the results of the two approaches is not large, we only used BinSpect-kmens for comparison. The same data pre-processing in Visium data was adopted. For BC23209_C1_stdata data, all genes were included for analysis. Then Benjamini-Yekutieli correction was performed to control FDR, the SE gene is regarded significant if the FDR was below the threshold of 0.05.

Data availability

Four publicly available spatial transcriptomic datasets are used in this paper, including two mouse olfactory bulb data and two human breast cancer data (https://www.spatialresearch.org/).

Code availability

SPRI is implemented in Python. All source code can be found in the supporting website: https://github.com/xiaoyeye/SPRI, and all the published data and code can be downloaded as described in the paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61725302, 62073219, 62103262).

References

1. Larsson, L., J. Frisén, and J. Lundeberg, *Spatially resolved transcriptomics adds a new dimension to genomics.* Nature Methods, 2021. **18**(1): p. 15-18.
2. Zhuang, X., *Spatially resolved single-cell genomics and transcriptomics by imaging.* Nature Methods, 2021. **18**(1): p. 18-22.
3. Eng, C.-H.L., et al., *Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+.* Nature, 2019. **568**(7751): p. 235-239.
4. Xia, C., et al., *Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression.* Proceedings of the National Academy of Sciences, 2019. **116**(39): p. 19490-19499.
5. He, B., et al., *Integrating spatial gene expression and breast tumour morphology via deep learning.* Nature biomedical engineering, 2020. **4**(8): p. 827-834.
6. Rodriques, S.G., et al., *Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution.* Science, 2019. **363**(6434): p. 1463-1467.
7. Ståhl, P.L., et al., *Visualization and analysis of gene expression in tissue sections by spatial transcriptomics.* Science, 2016. **353**(6294): p. 78-82.
8. Berglund, E., et al., *Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity.* Nature communications, 2018. **9**(1): p. 1-13.
9. Ji, A.L., et al., *Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma.* Cell, 2020. **182**(2): p. 497-514. e22.
10. Edsgärd, D., P. Johnsson, and R. Sandberg, *Identification of spatial expression trends in single-cell gene expression data.* Nature methods, 2018. **15**(5): p. 339-342.
11. Svensson, V., S.A. Teichmann, and O. Stegle, *SpatialDE: identification of spatially variable genes.* Nature methods, 2018. **15**(5): p. 343-346.
12. Miller, B.F., et al., *Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities.* Genome research, 2021. **31**(10): p. 1843-1855.
13. Dries, R., et al., *Giotto: a toolbox for integrative analysis and visualization of spatial expression data.* Genome biology, 2021. **22**(1): p. 1-31.
14. Sun, S., J. Zhu, and X. Zhou, *Statistical analysis of spatial expression patterns for spatially*

resolved transcriptomic studies. Nature methods, 2020. **17**(2): p. 193-200.

15. Hu, J., et al., *SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network.* Nature methods, 2021: p. 1-10.
16. Reshef, D.N., et al., *Detecting novel associations in large data sets.* science, 2011. **334**(6062): p. 1518-1524.
17. Reshef, Y.A., et al., *Measuring dependence powerfully and equitably.* The Journal of Machine Learning Research, 2016. **17**(1): p. 7406-7468.
18. Tepe, B., et al., *Single-cell RNA-seq of mouse olfactory bulb reveals cellular heterogeneity and activity-dependent molecular census of adult-born neurons.* Cell reports, 2018. **25**(10): p. 2689-2703. e3.
19. Zheng, R., et al., *Transcriptomic insights into the response of the olfactory bulb to selenium treatment in a mouse model of Alzheimer' s disease.* International journal of molecular sciences, 2019. **20**(12): p. 2998.
20. Nunes, D. and T. Kuner, *Disinhibition of olfactory bulb granule cells accelerates odour discrimination in mice.* Nature communications, 2015. **6**(1): p. 1-13.
21. Mi, W., et al., *Cystatin C inhibits amyloid-β deposition in Alzheimer's disease mouse models.* Nature genetics, 2007. **39**(12): p. 1440-1442.
22. Tian, Y., et al., *FTH1 inhibits ferroptosis through ferritinophagy in the 6-OHDA model of Parkinson' s disease.* Neurotherapeutics, 2020. **17**(4): p. 1796-1812.
23. Guo, X., et al., *Oxidative damage to the TCA cycle enzyme MDH1 dysregulates bioenergetic enzymatic activity in the aged murine brain.* Journal of proteome research, 2020. **19**(4): p. 1706-1717.
24. Haider, S., et al., *Spirulina platensis reduces the schizophrenic-like symptoms in rat model by restoring altered APO-E and RTN-4 protein expression in prefrontal cortex.* Life Sciences, 2021. **277**: p. 119417.
25. Wang, T., et al., *LINC01116 promotes tumor proliferation and neutrophil recruitment via DDX5-mediated regulation of IL-1β in glioma cell.* Cell death & disease, 2020. **11**(5): p. 1-13.
26. Cuvertino, S., et al., *ACTB loss-of-function mutations result in a pleiotropic developmental disorder.* The American Journal of Human Genetics, 2017. **101**(6): p. 1021-1033.
27. Aiken, J., et al., *Tubulin mutations in brain development disorders: Why haploinsufficiency does not explain TUBA1A tubulinopathies.* Cytoskeleton, 2020. **77**(3-4): p. 40-54.
28. Perucho, L., et al., *RPLP1, a crucial ribosomal protein for embryonic development of the nervous system.* PloS one, 2014. **9**(6): p. e99956.
29. Lever, J., et al., *CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer.* Nature methods, 2019. **16**(6): p. 505-507.
30. Gu, Y., et al., *A pan-cancer analysis of the prognostic and immunological role of β-actin (ACTB) in human cancers.* Bioengineered, 2021. **12**(1): p. 6166-6185.
31. Zhang, X., et al., *Thymosin beta 10 is a key regulator of tumorigenesis and metastasis and a novel serum marker in breast cancer.* Breast Cancer Research, 2017. **19**(1): p. 1-15.
32. Dong, H., et al., *Long non-coding RNA SNHG14 induces trastuzumab resistance of breast cancer via regulating PABPC1 expression through H3K27 acetylation.* Journal of cellular and molecular medicine, 2018. **22**(10): p. 4935-4947.

33. Luo, Y., et al., *Loss of ASAP3 destabilizes cytoskeletal protein ACTG1 to suppress cancer cell migration.* Molecular medicine reports, 2014. **9**(2): p. 387-394.
34. Deng, Y., et al., *Correlations Between the Characteristics of Alternative Splicing Events, Prognosis, and the Immune Microenvironment in Breast Cancer.* Frontiers in Genetics, 2021. **12**: p. 973.
35. Saidy, B., et al., *PP1, PKA and DARPP-32 in breast cancer: A retrospective assessment of protein and mRNA expression.* Journal of Cellular and Molecular Medicine, 2021.
36. Shen, L., et al., *Metabolic reprogramming in triple-negative breast cancer through Myc suppression of TXNIP.* Proceedings of the National Academy of Sciences, 2015. **112**(17): p. 5425-5430.
37. Marinov, G.K., et al., *From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing.* Genome research, 2014. **24**(3): p. 496-510.
38. Jin, S., et al., *Inference and analysis of cell-cell communication using CellChat.* Nature communications, 2021. **12**(1): p. 1-20.

Figures

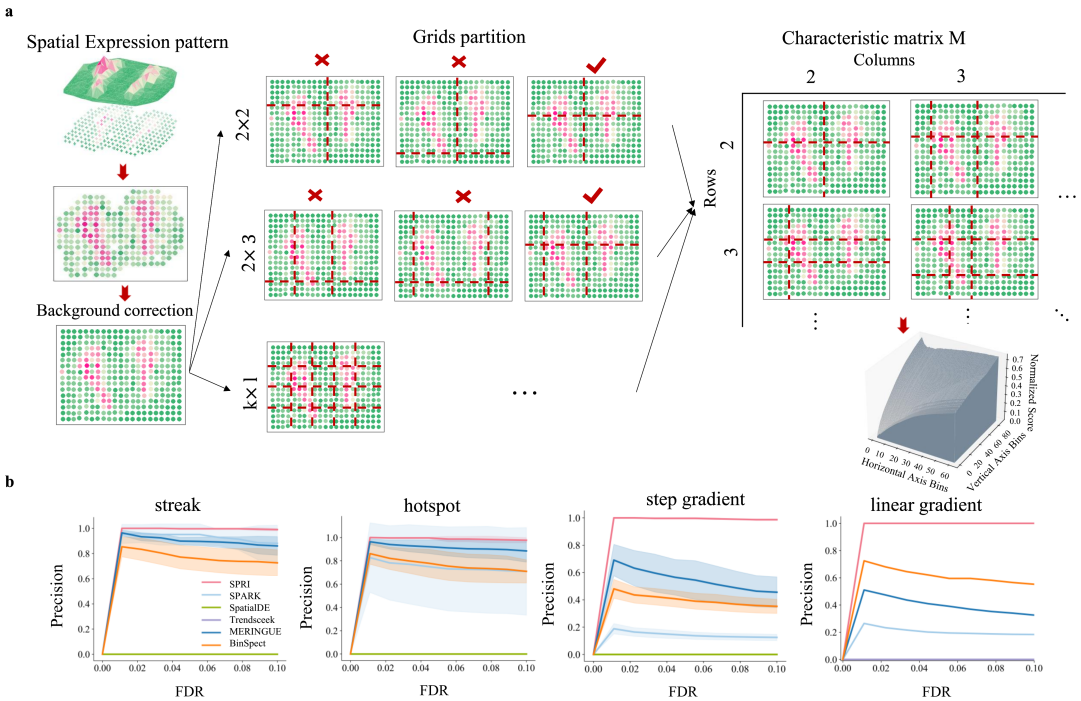

Fig.1 | Overview of the SPRI method and results on the simulation dataset. a, Overview of SPRI algorithm. SPRI converts the spatial gene pattern problem to association detection problem between coordinates values of (x, y) with gene read count as the observed frequencies and then calculates their total information coefficient (TIC) using all possible x-y-grids. **b**, Precision plots for all four different simulation patterns, including Hotspot, Streak, Step gradient and Linear gradient, which display the proportion of true positive samples among retrieved genes (Y-axis) detected by the compared methods at different FDRs (X-axis).

Fig.2 | Results of the mouse olfactory bulb data (MOB Replicate 11). a, Power plot that displays the number of genes with SE patterns (*Y*-axis) detected by six different methods, i.e., SPRI (pink), SPARK (light blue), SpatialDE (green), Trendsceek (purple), MERINGUE (dark blue) and BinSpect (orange) at different FDRs (*X*-axis), respectively. **b,** Enrichment of SE genes that are verified in the study of Tepe et al. [18]. SE genes are defined uisng a 0.05 FDR cutoff and *P* value is calculated by Fisher's exact test. **c,** Percentage (pink/blue) of SPRI/SPARK SE genes that are verified in SPARK/SPRI top-ranked SE genes. **d,** Boxplot of the expression levels of SE genes uniquely identified by SPRI, SPARK, SpatialDE, MERINGUE, BinSpect and union set of the five. Each gray point represents the average expression of an SE gene across all spots. **e**, Visualization of gene spatial expression patterns from MOB Replicate 11 dataset for the two genes only detected by SPRI (*Scg2* and *Gabrb3*) plus four genes that are detected by both SPRI and SPARK. The color represents relative expression level of

the gene (pink: high; green: low). **f**, *In situ* hybridization obtained from the Mouse Brain Atlas of the Allen Brain Atlas for three representative genes (*Scg2*, *Ptgds*, and *Gabrb3*). **g**, GO enrichment analysis on top 100 SE genes by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect respectively. **h**, Bubble plot of enriched GO terms and KEGG pathways (purple) on the whole SE genes at FDR cutoff of 0.05 by SPRI. GO term is colored according to three different categories: Biological Processes (blue), Cellular Components (green), and Molecular Functions (yellow).

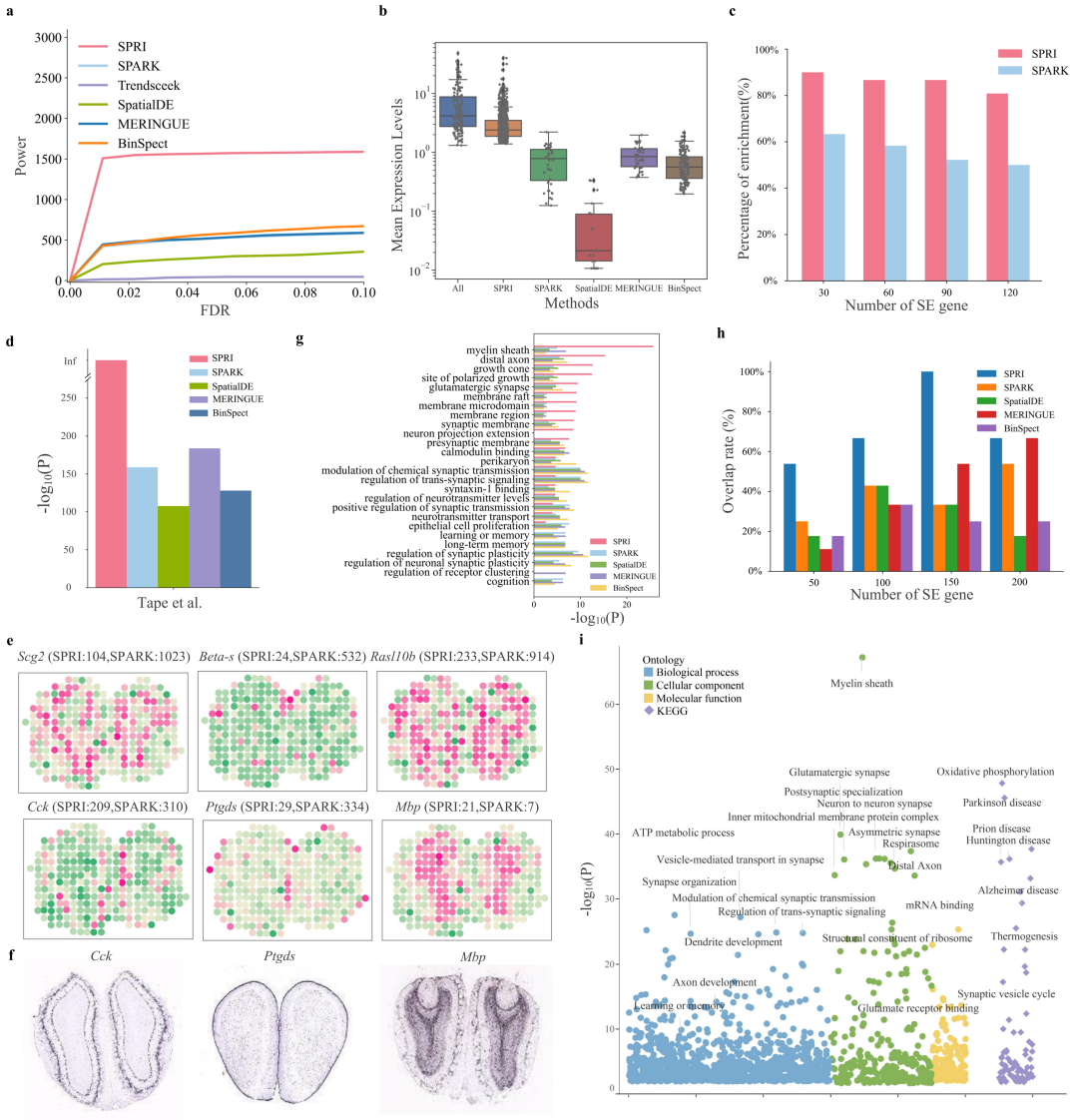

Fig.3 | Results of the mouse olfactory bulb data (MOB Replicate 12). a, Power plot that displays the number of genes with SE patterns (*Y*-axis) detected by six different methods respectively. **b,** Expression levels of SE genes uniquely identified by SPRI, SPARK, SpatialDE, MERINGUE, BinSpect and union set of the five. **c,** Percentage (pink/blue) of SPRI/SPARK SE genes that are verified in SPARK/SPRI top-ranked SE genes. **d**, Enrichment of SE genes that are verified in the study of Tepe et al.[18]. **e**, Visualization of gene spatial expression patterns from MOB Replicate 12 dataset for representative genes. **f**, *In situ* hybridization obtained from the Mouse Brain Atlas of the Allen Brain Atlas for three genes (*Cck*, *Ptgds*, and *Mbp*). **g**, GO enrichment analysis on top 100 SE genes by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect respectively. **h**, Overlap rate of the top 10 GO terms detected by SPRI, SPARK,

SpatialDE, MERINGUE and BinSpect between MOB Replicate 11 and MOB Replicate 12 data for the same number of genes. **i**, Bubble plot of enriched GO terms and KEGG pathways (purple) on the whole SPRI SE genes.

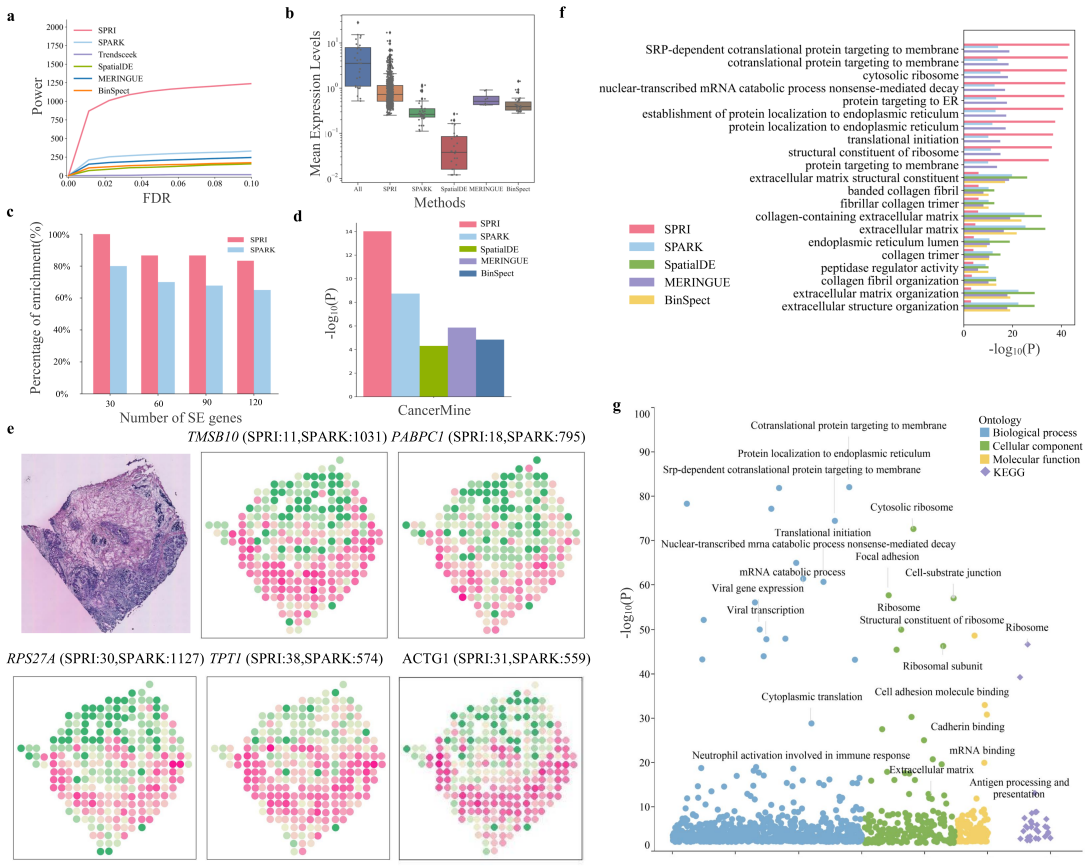

Fig.4 | Results of the human breast cancer data (Breast Cancer Layer 2). a, Power plot that displays the number of genes with SE patterns (*Y*-axis) detected by six different methods at different FDRs (*X*-axis), respectively. **b**, Boxplot of expression levels of SE genes identified by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect in Breast Cancer Layer 2 data. **c,** Percentage (pink/blue) of SPRI/SPARK SE genes that are verified in SPARK/SPRI top-ranked SE genes. **d**, Enrichment of SE genes that are verified in the study of CancerMine database. **e**, Visualization of gene spatial expression patterns for five genes that are detected by SPRI only. The first one is hematoxylin & eosin stained brightfield image of Breast Cancer Layer 2 from ref. [7]. **f**, GO enrichment analysis on top 100 SE genes by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect respectively. **g**, Bubble plot of enriched GO terms and KEGG pathways (purple) on the whole SPRI SE genes at FDR = 0.05.

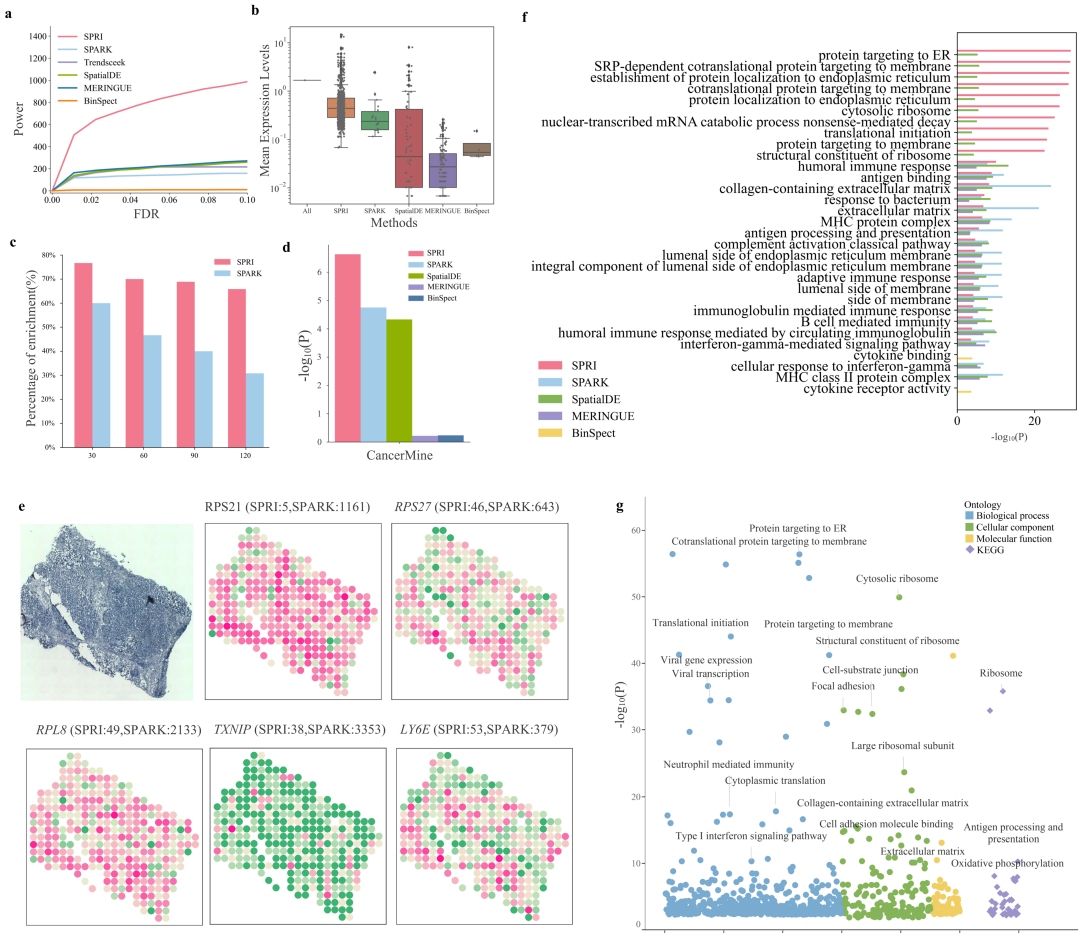

Fig.5 | Results of the human breast cancer data (BC23209_C1_stdata). a, Power plot that displays the number of genes with SE patterns (*Y*-axis) detected by six different methods, i.e., SPRI, SPARK, SpatialDE, Trendsceek, MERINGUE and BinSpect, at different FDRs (*X*-axis), respectively. **b**, Boxplot shows the expression levels of SE genes identified by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect in BC23209_C1_stdata. **c,** Percentage (pink/blue) of SPRI/SPARK SE genes that are verified in SPARK/SPRI top-ranked SE genes. **d**, Enrichment of SE genes that are verified in the study of CancerMine database. **e**, Visualization of gene spatial expression patterns for five genes that are identified by SPRI only. The first one is hematoxylin & eosin stained brightfield image of BC23209_C1_stdata from ref. [5]. **f**, GO enrichment analysis on top 100 SE genes by SPRI, SPARK, SpatialDE, MERINGUE and BinSpect respectively. **g**, Bubble plot of enriched GO terms and KEGG pathways (purple) on the whole SPRI SE genes.