

Supplementary of HEARTSVG

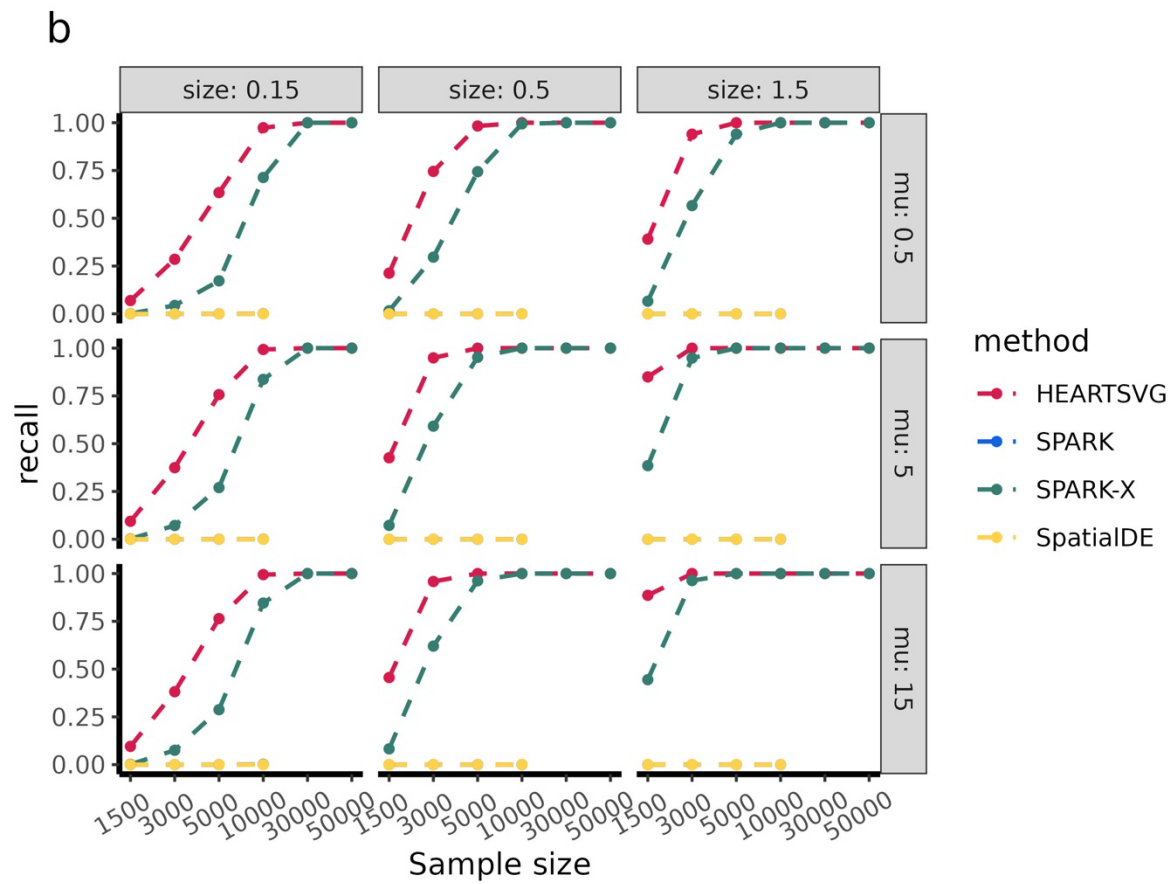
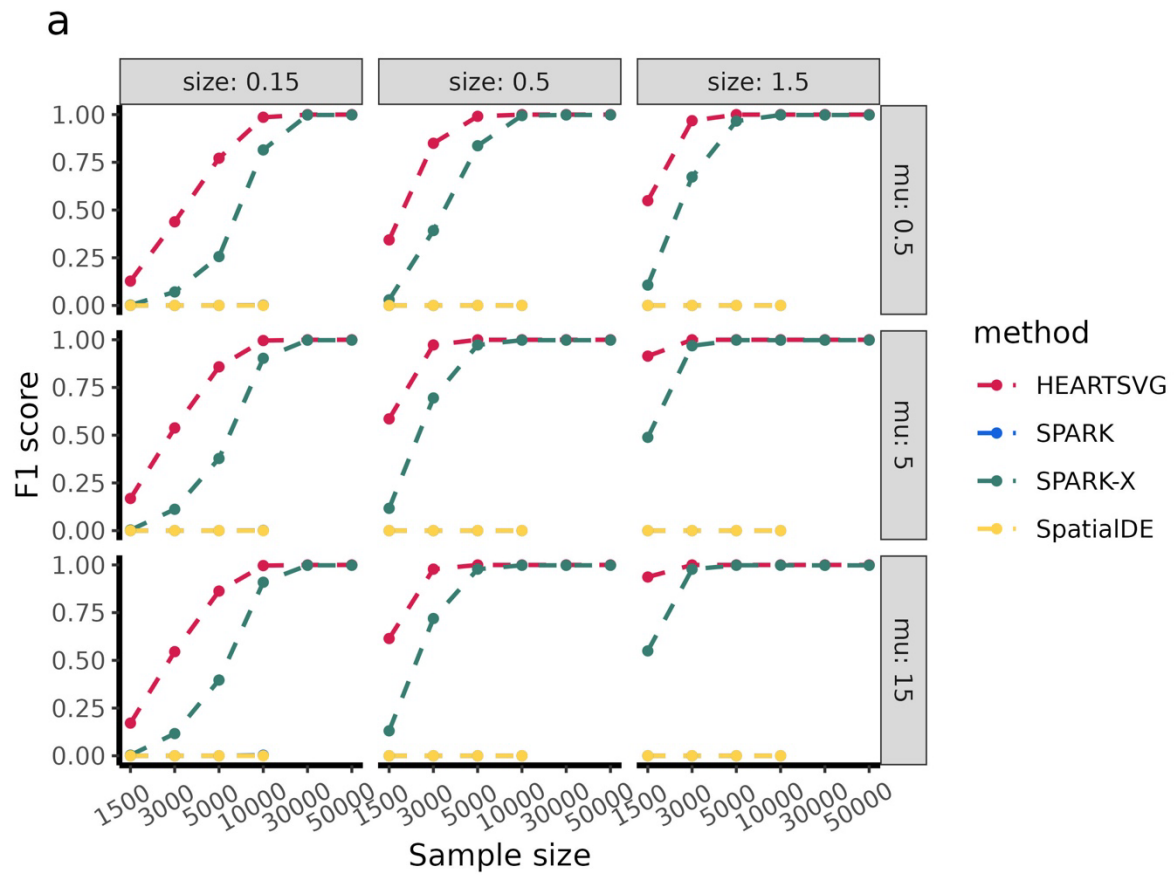
Xin Yuan, Yanran Ma, Ruitian Gao, Shuya Cui, Yifan Wang, Botao Fa, Shiyang Ma, Shuangge Ma, and
Zhangsheng Yu

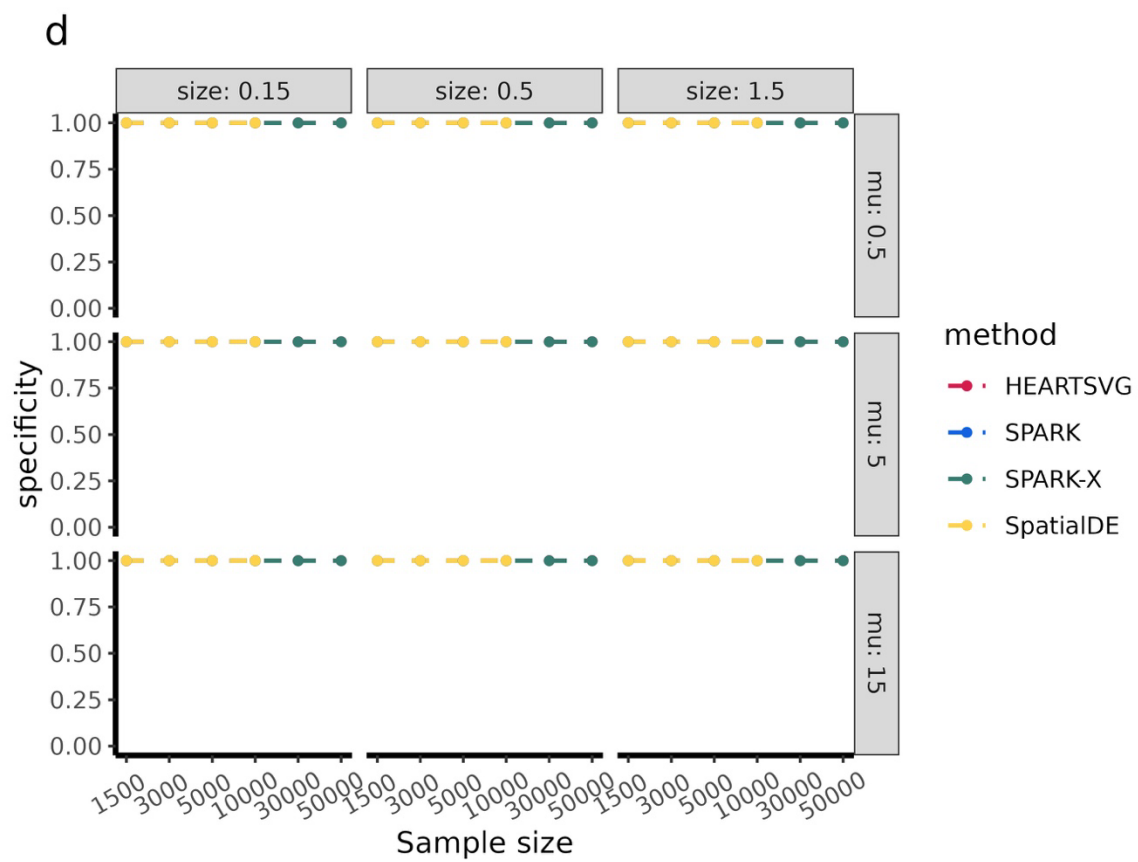
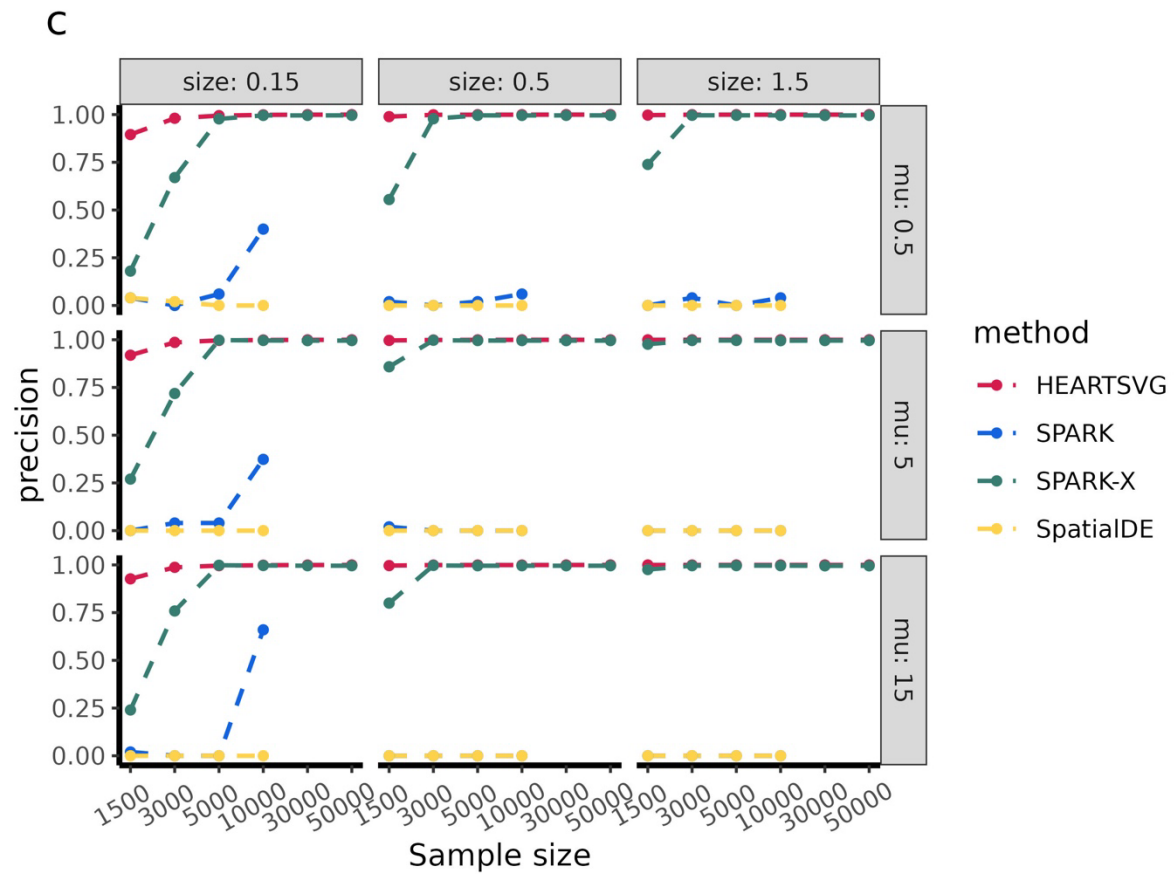
1. Simulation Results

1.1 Simulation 1 additional figures

We used F_1 score, recall, precision, specificity, and false positive (FP) to comprehensively evaluate the performance of HEARTSVG and SpatialDE, SPARK, and SPARK-X. All simulation settings were illustrated in the two sections, Simulation and Methods. Each simulation scenario has 50 replications. Simulation datasets were generated by varying the sample size (from 1500 to 50000), ZINB parameters ($size = 0.15, 0.5, 1.5, mu = 0.5, 5, 15$), and SVG percentages (hotspot, streak =5%, gradient=15%). The paper presented the simulation scenario with ZINB parameters ($size = 0.5, mu = 0.5$). The complete simulation results were shown in this supplementary.

All simulation results of the hotspot pattern





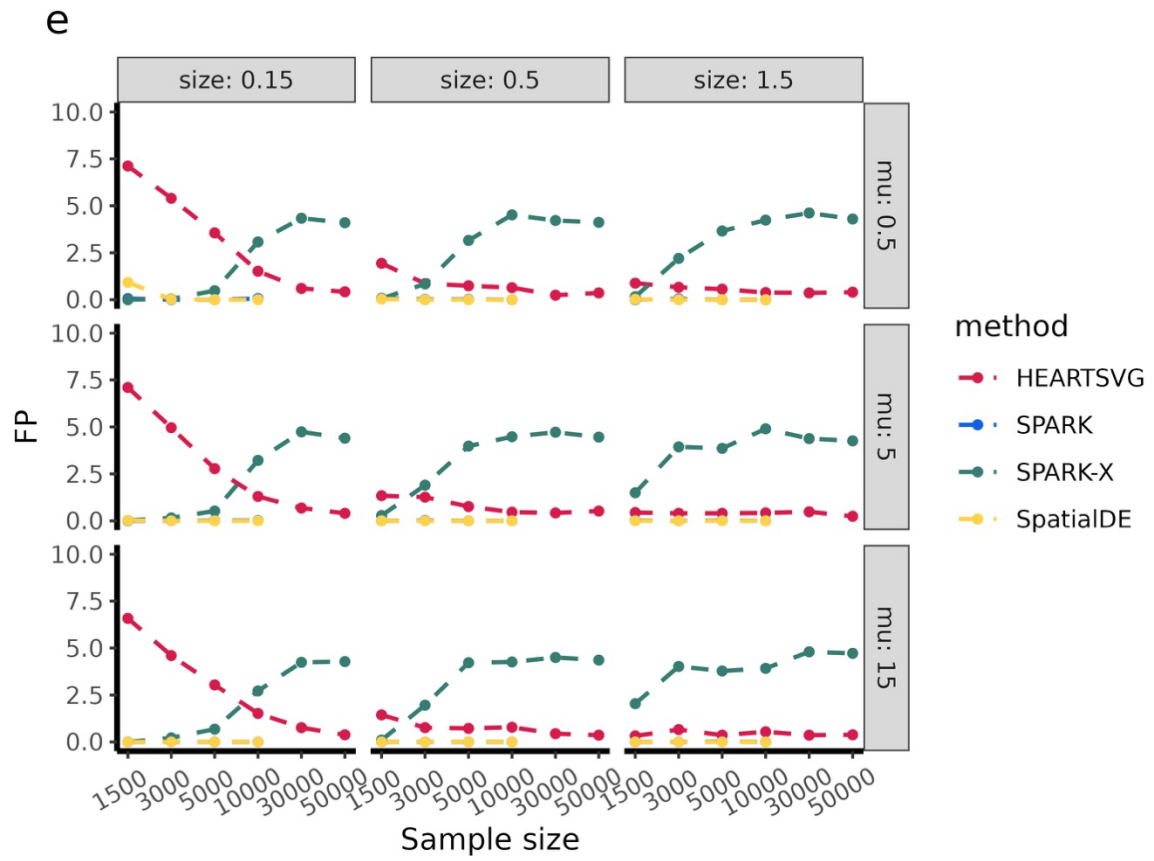
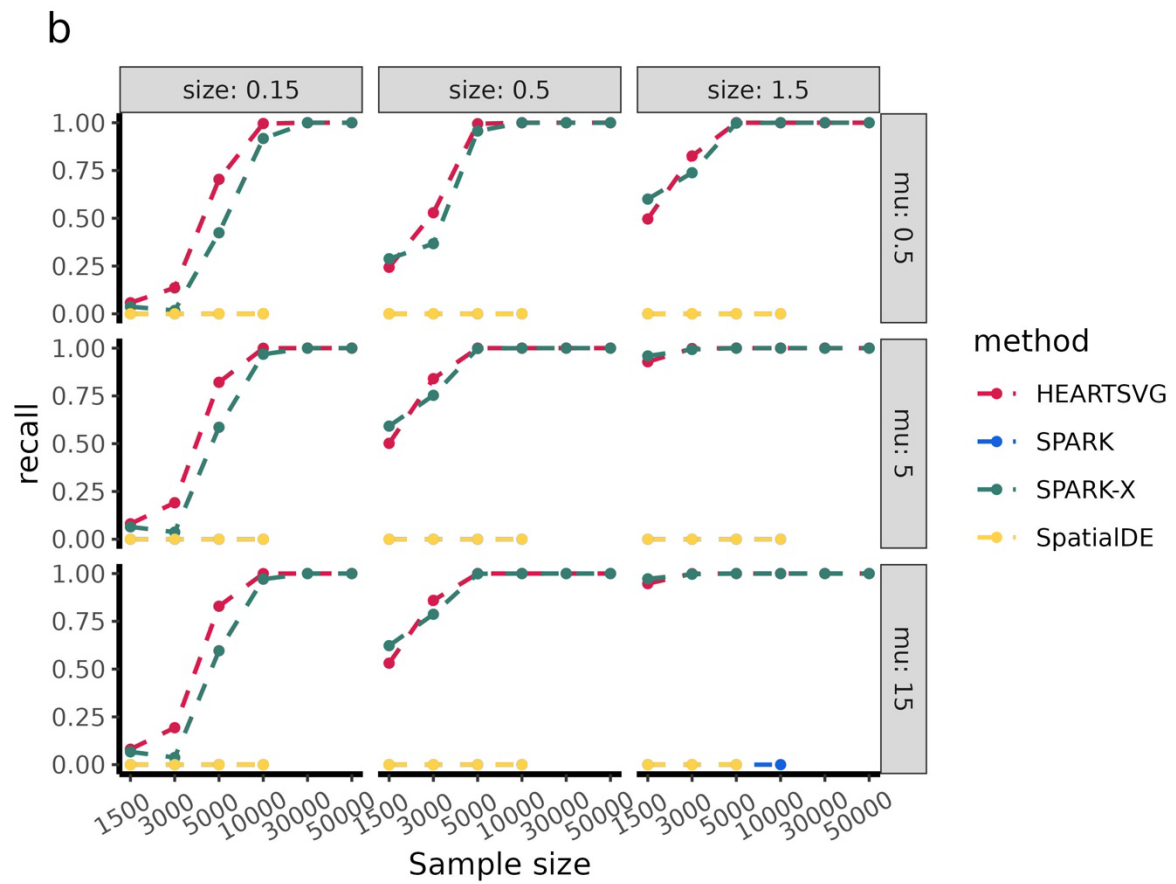
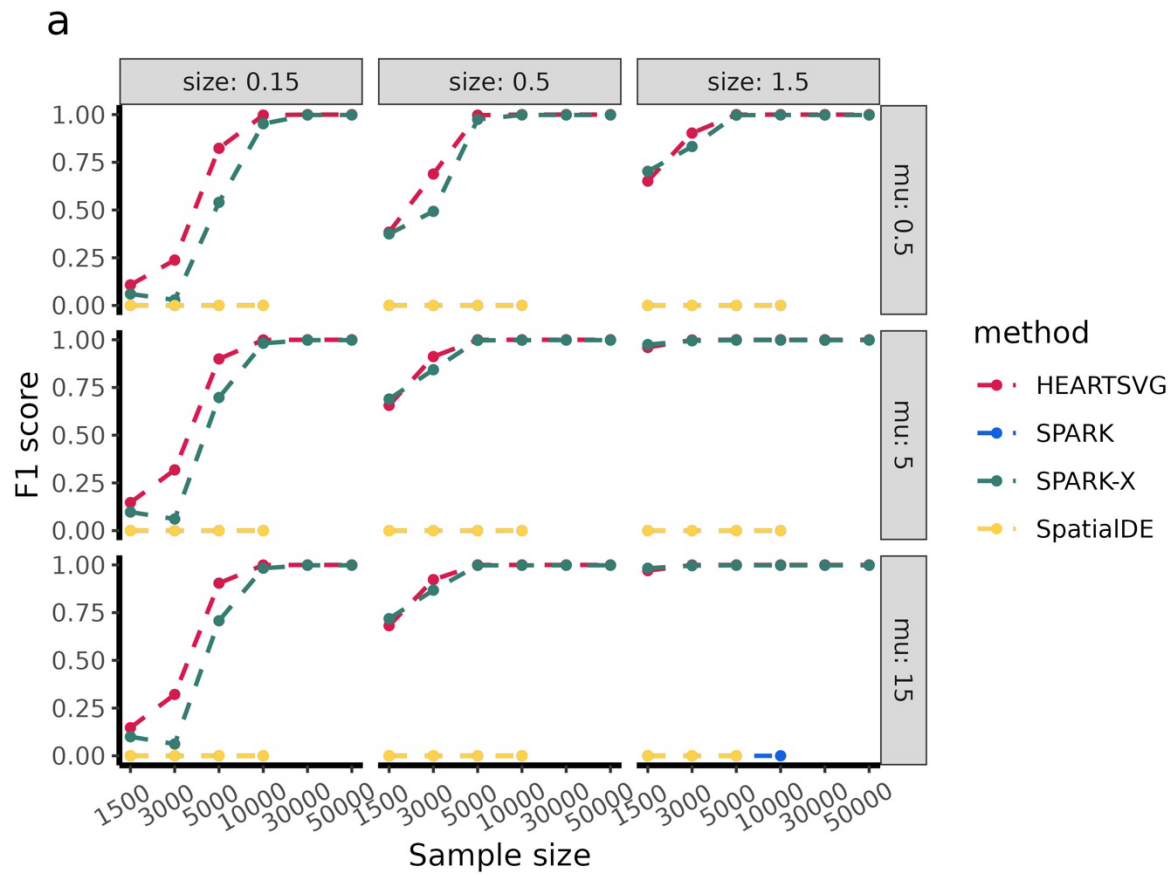
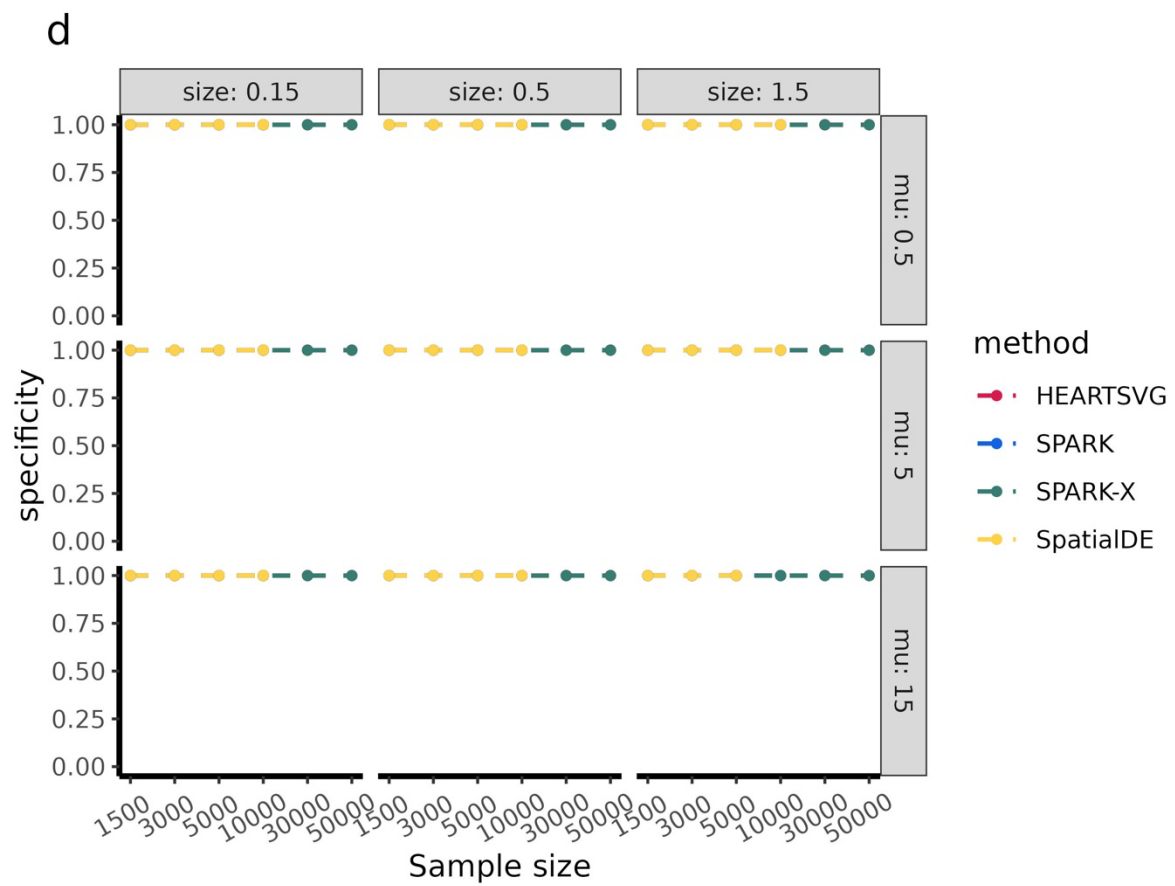
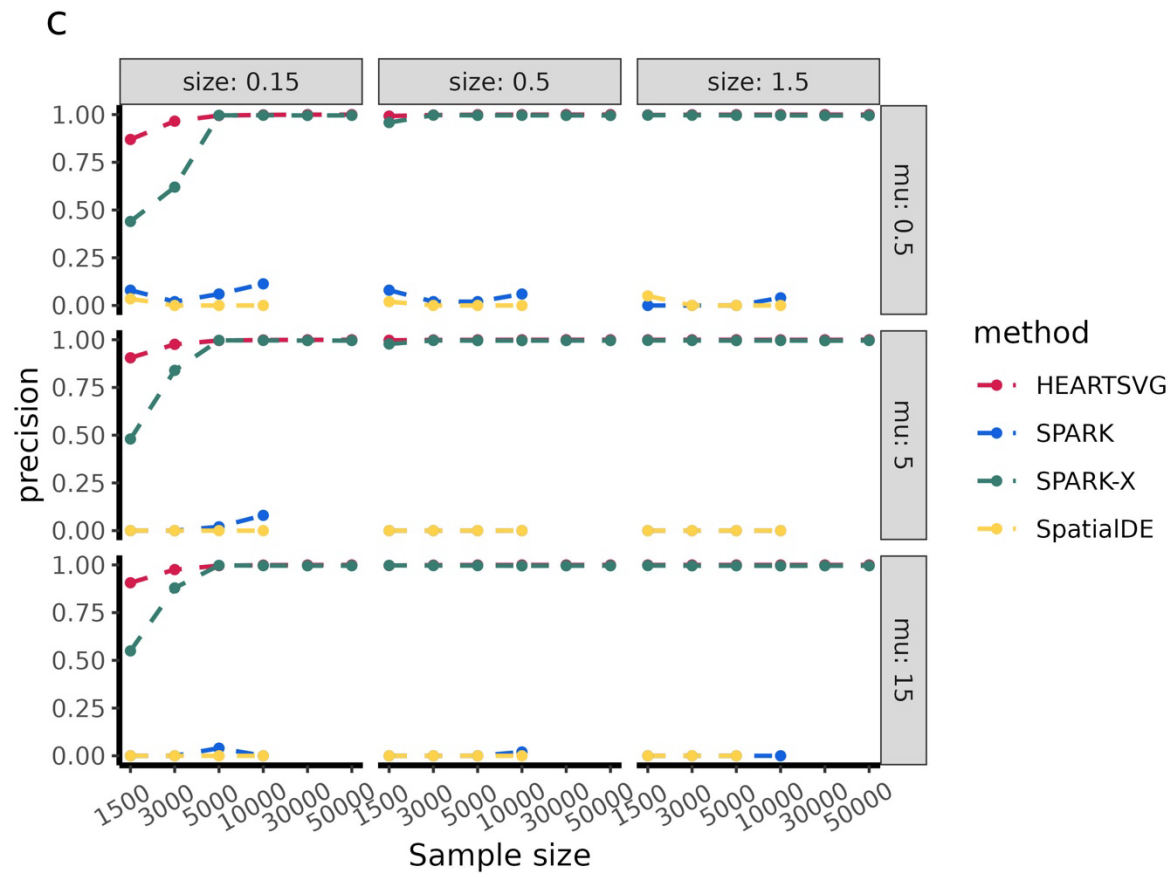


Figure S1 Simulation results of hotspot pattern. (a) F1 score, (b) recall, (c) precision, (d) specificity, (e) FP.

All simulation results of streak pattern





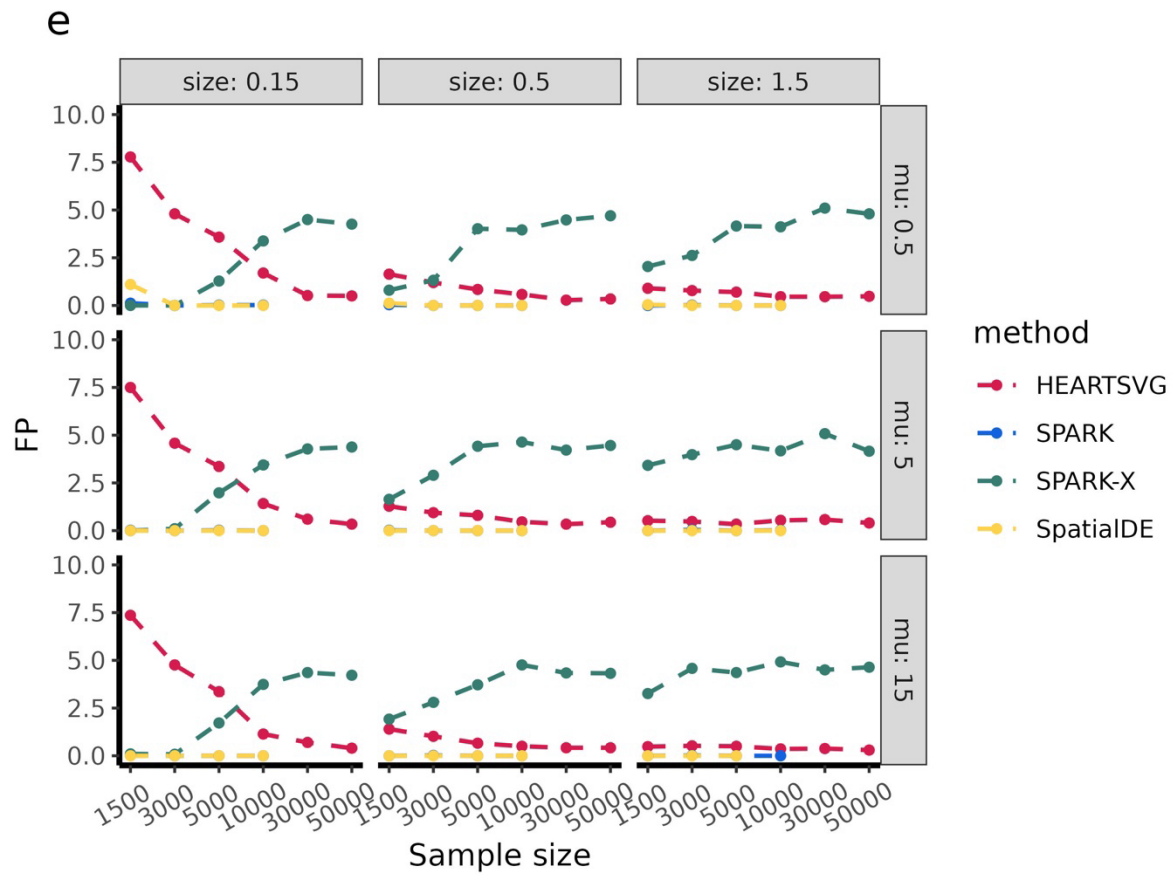
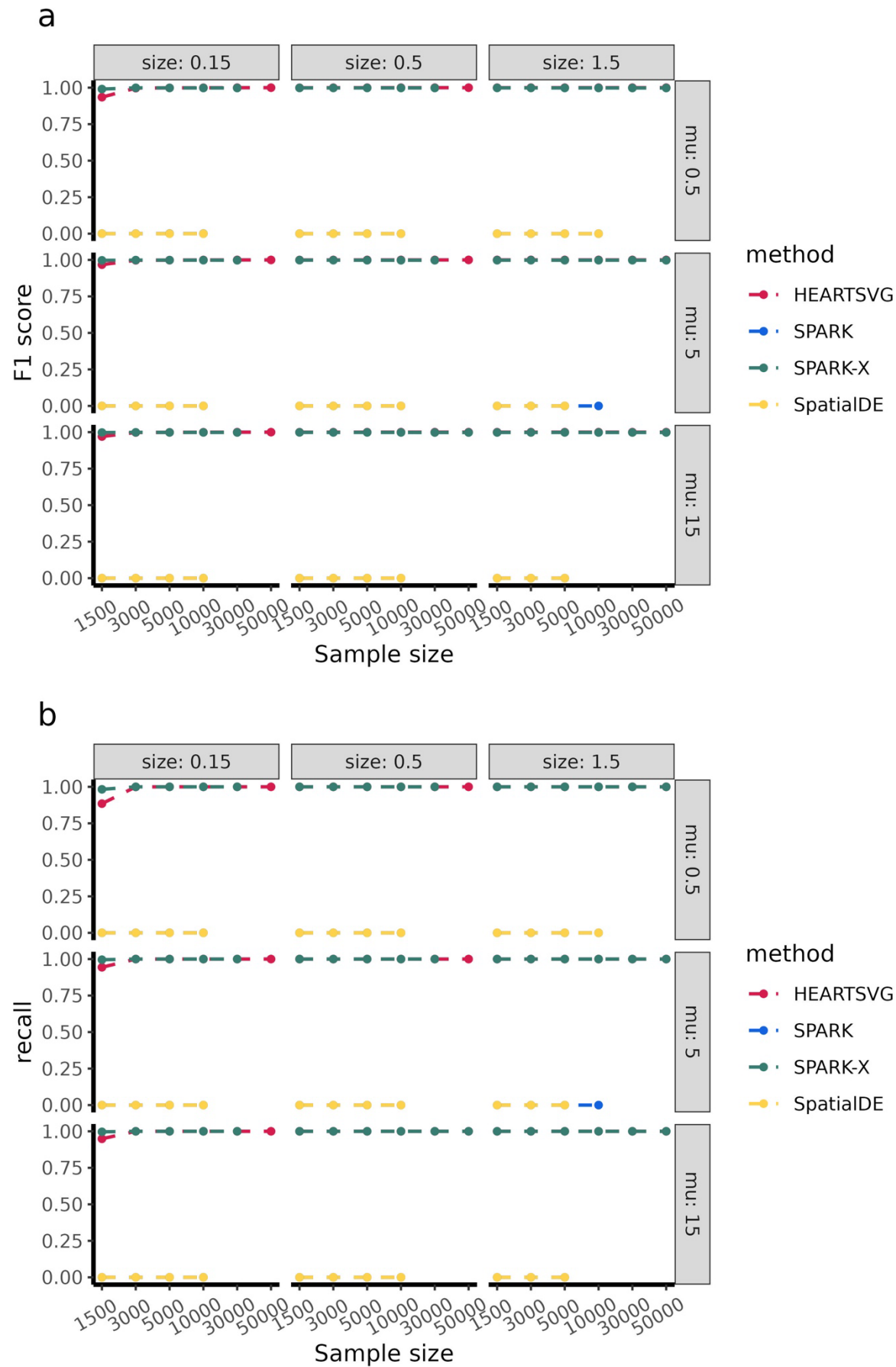
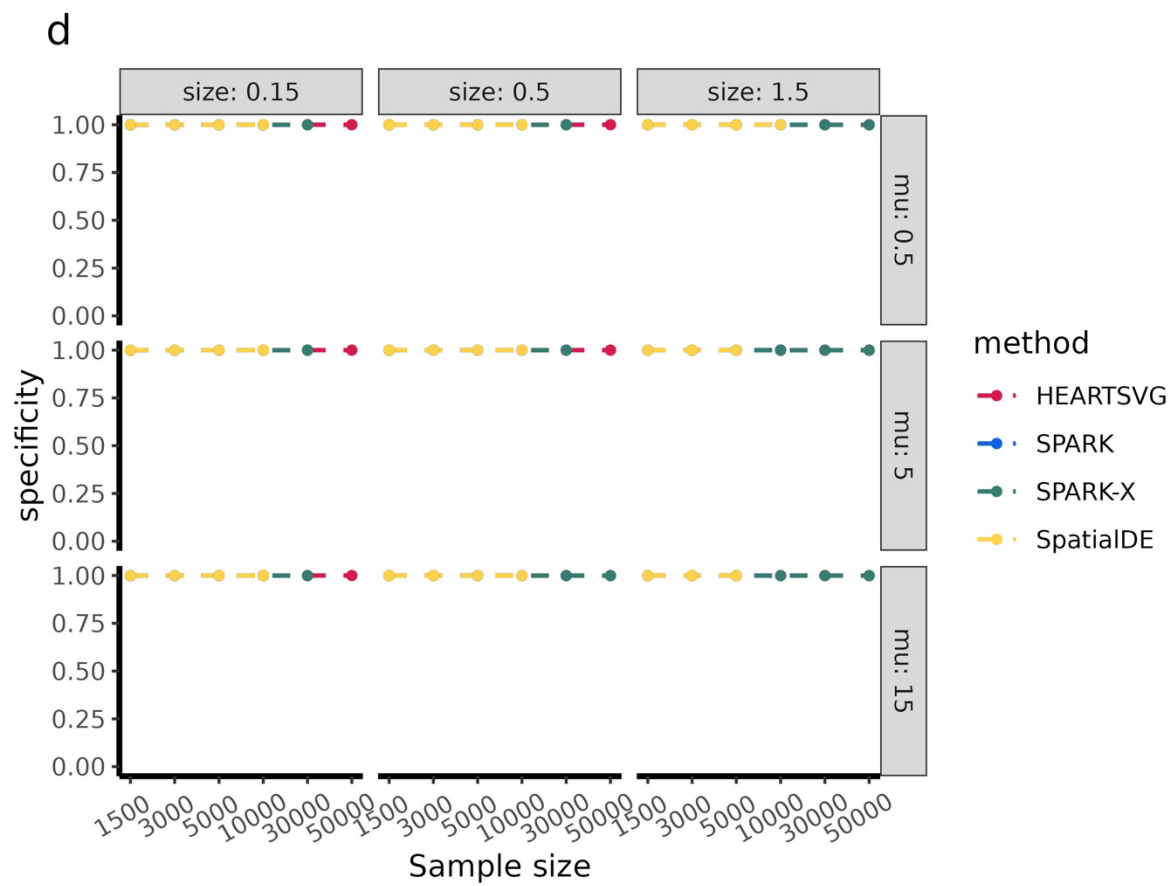
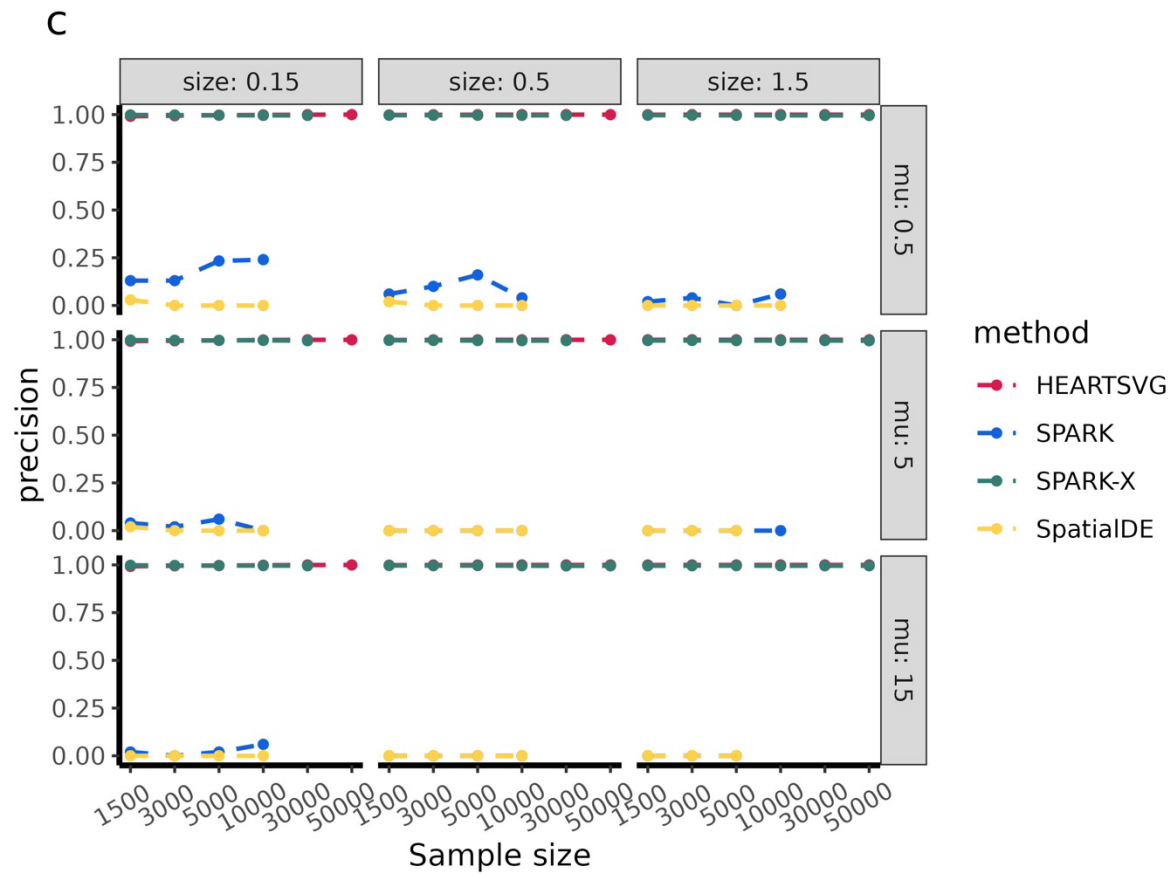


Figure S2 Simulation results of streak pattern. (a) F1 score, (b) recall, (c) precision, (d) specificity, (e) FP.

All simulation results of gradient pattern





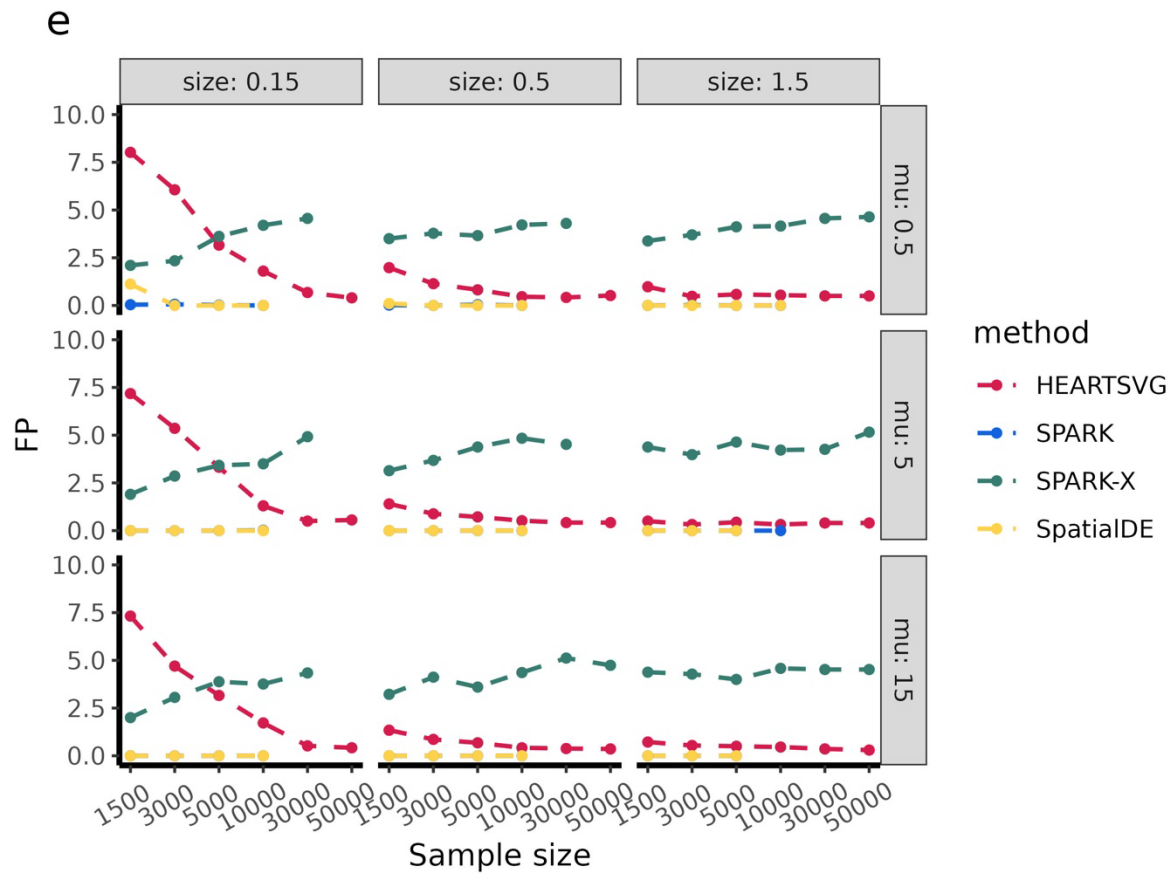


Figure S3 Simulation results of gradient pattern. (a) F1 score, (b) recall, (c) precision, (d) specificity, (e) FP.

1.2 Simulation 2 additional figures

Because the uncertainty surrounding the number of SVGs in real data is unknown, we generated simulation datasets with varying percentages (from 0% to 50%) of SVGs and sample sizes (from 3000 to 10,000) in three representative spatial patterns. We also compared the performance of HEARTSVG and SPARK-X using F_1 score, and false positive (FP). The paper presented the simulation scenario with ZINB parameters ($size = 0.5, mu = 0.5$) and moderate sample size ($n=5000$). The complete simulation results were shown in this supplementary.

Hotspot pattern with varying percentages of SVGs

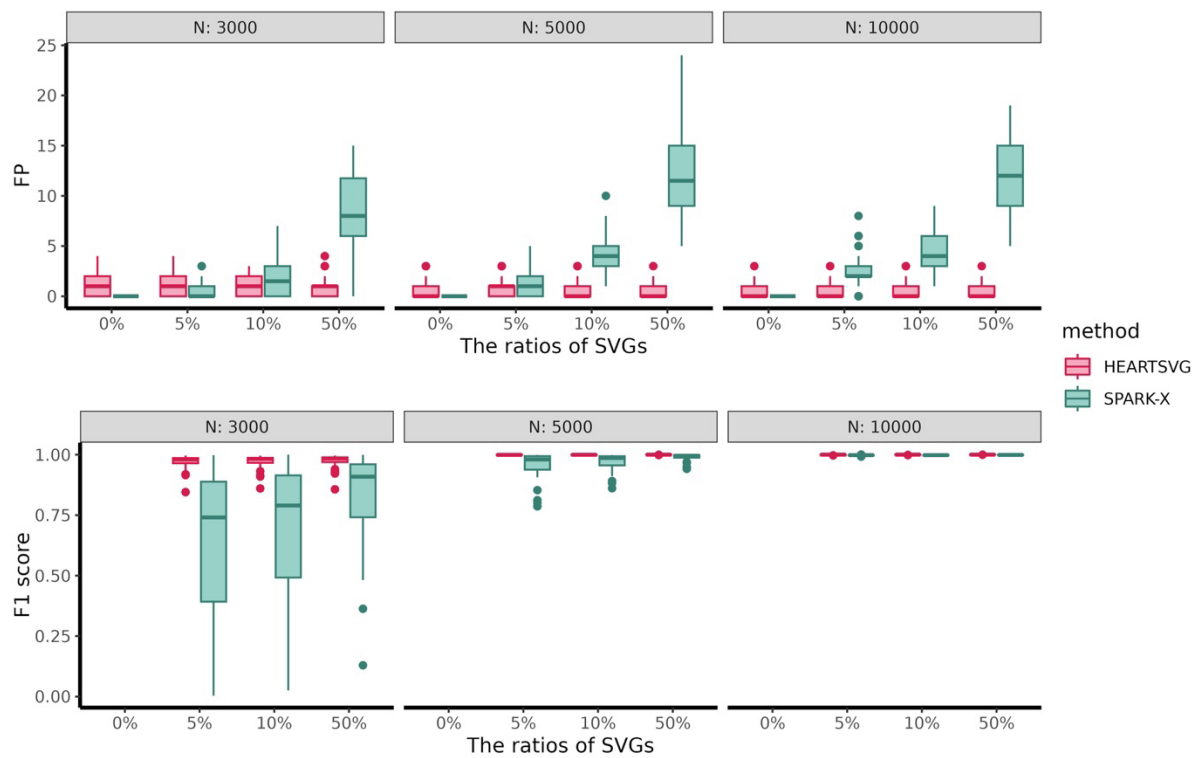


Figure S4 Simulation2 results of hotspot pattern, F1 score and FP.

Streak pattern with varying percentages of SVGs

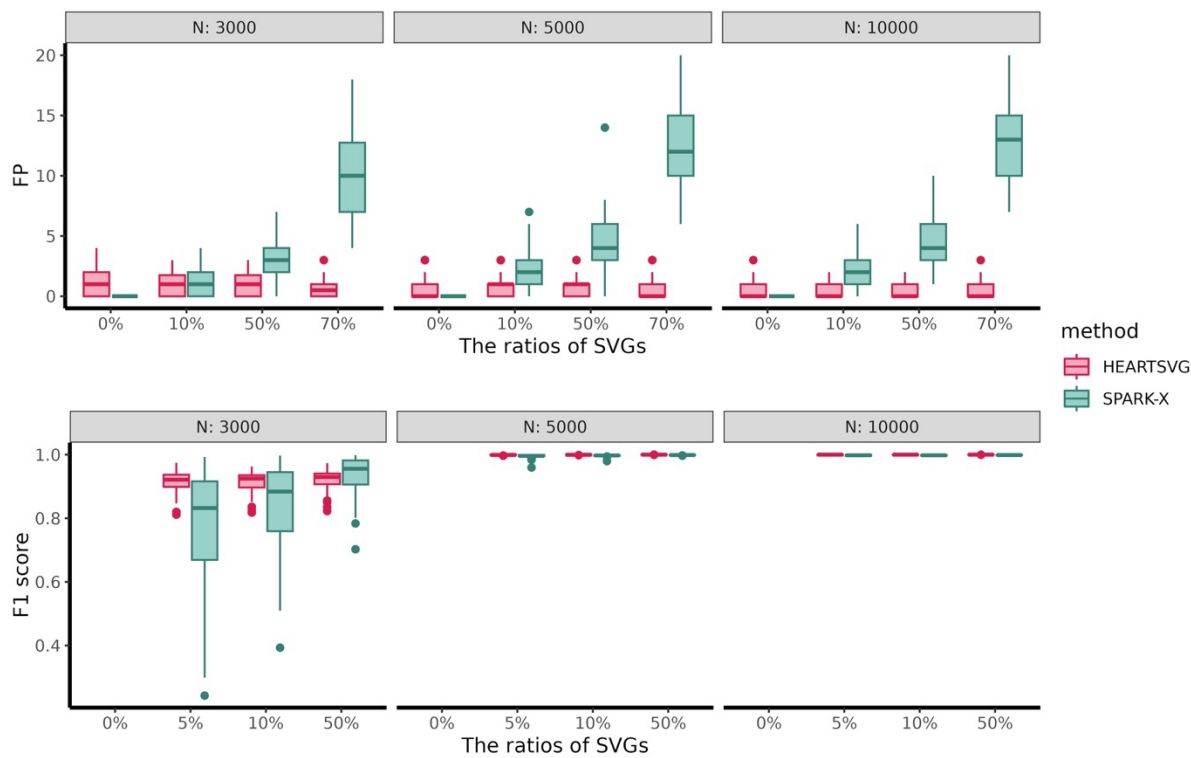


Figure S5 Simulation2 results of streak pattern, F1 score and FP.

Gradient pattern with varying percentages of SVGs

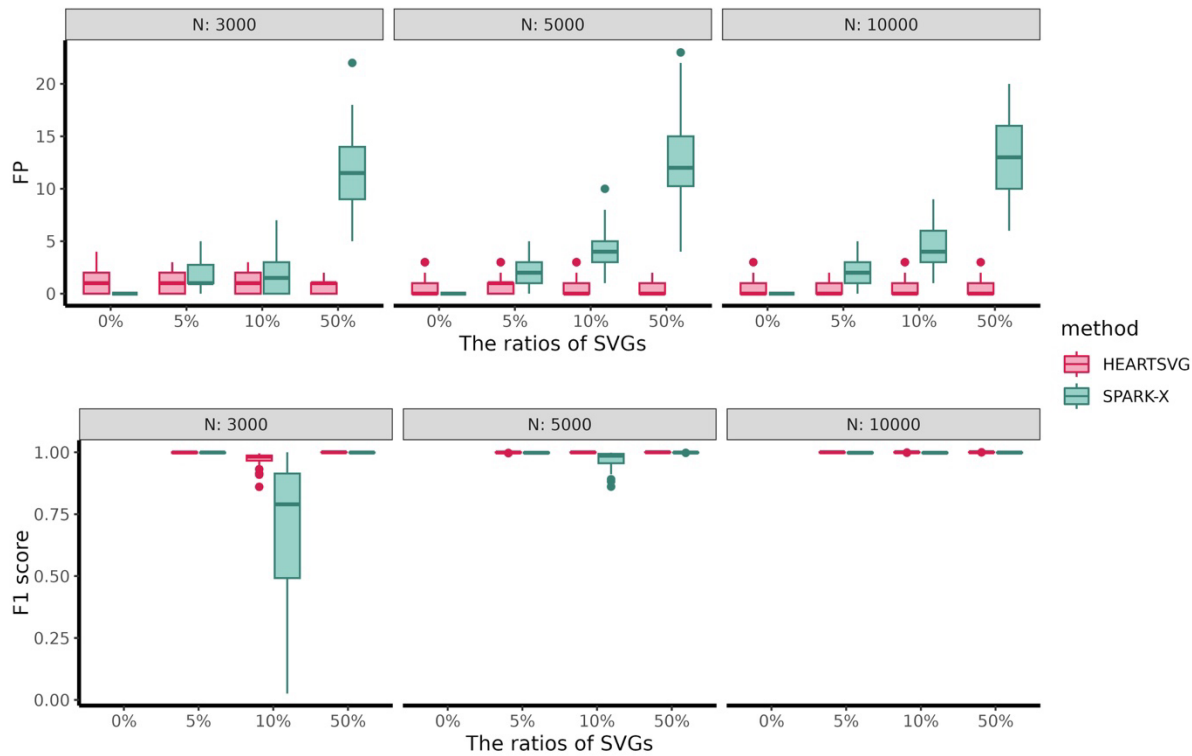


Figure S6 Simulation2 results of gradient pattern, F1 score and FP.

2. Methods

$e_{x_1 y_1}$	$e_{x_1 y_2}$	...	$e_{x_1 y_{n_2}}$
$e_{x_2 y_1}$	$e_{x_2 y_2}$	...	$e_{x_2 y_{n_2}}$
		$\vdots$	
$e_{x_{n_1} y_1}$	$e_{x_{n_1} y_2}$	...	$e_{x_{n_1} y_{n_2}}$

Figure S7 Example of ST data.

2.1 The following demonstrates the rationality of HEARTSVG.

For a gene g without the spatial pattern in the ST data, its expression count e is independent of its coordinates (x, y) . Define $f(x, y, e) = g(e|\boldsymbol{\theta})$, the gene g 's expression e of each spot is only dependent on the parameters $\boldsymbol{\theta}$ (e.g., $\boldsymbol{\theta}$ could be the λ of the Poisson distribution, (α, β) of the Gamma distribution, etc.).

That is, the gene g 's expression counts e of the SRT data are independent and identically distributed random variables (Abbreviated as iid random variables).

We assume that every row has n_2 elements, every column has n_1 elements and expression matrix are a $n_1 * n_2$ matrix (Fig.S8). Next, take the marginal expression series obtained by the semi-pooling process with the row direction and feature map $(1 * n_2)$ as an example.

The gene g 's expressions of the first row are denoted by $r_{x_1} = (e_{x_1 y_1}, \dots, e_{x_1 y_j}, \dots, e_{x_1 y_{n_2}})^T$.

The joint distribution of the first row is $f(r_{x_1}|\boldsymbol{\theta}) = f(e_{x_1 y_1}, \dots, e_{x_1 y_j}, \dots, e_{x_1 y_{n_2}}|\boldsymbol{\theta}) = \prod_{j=1}^{n_2} g(e_{x_1 y_j}|\boldsymbol{\theta})$. Similarly, the expression vector and joint distribution of the second row

are $r_{x_2} = (e_{x_2 y_1}, \dots, e_{x_2 y_j}, \dots, e_{x_2 y_{n_2}})^T$ and $f(e_{x_2 y_1}, \dots, e_{x_2 y_j}, \dots, e_{x_2 y_{n_2}}|\boldsymbol{\theta}) =$

$\prod_{j=1}^{n_2} g(e_{x_2 y_j}|\boldsymbol{\theta})$, respectively. The expression vector and joint distribution of the i -th row

are $r_{x_i} = (e_{x_i y_1}, \dots, e_{x_i y_j}, \dots, e_{x_i y_{n_2}})^T$ and $f(e_{x_i y_1}, \dots, e_{x_i y_j}, \dots, e_{x_i y_{n_2}} | \boldsymbol{\theta}) = \prod_{j=1}^{n_2} g(e_{x_i y_j} | \boldsymbol{\theta})$, respectively. Obviously, the joint distributions of each row's expressions are identically distributed. r_{x_1} and r_{x_2} have no overlapping elements. Hence, it is easy to prove that $f(r_{x_1}, r_{x_2} | \boldsymbol{\theta}) = f(r_{x_1} | \boldsymbol{\theta}) * f(r_{x_2} | \boldsymbol{\theta})$, which means r_{x_1} and r_{x_2} are independent and identically distributed random variables.

Similarly, we can prove that, $\forall i_1, i_2 = 1, \dots, n_1$, and $i_1 \neq i_2$, $r_{x_{i_1}}$ and $r_{x_{i_2}}$ are iid variables. By the similar derivation process as above, we can prove that the elements of the marginal expression series obtained by semi-pooling parameters with different parameters are iid variables. This is a very strong condition and is hard to verify empirically.

In practice, we assume that the expression counts of the non-SVG gene at a given location (x_i, y_j) are independent of expressions at nearby locations. Therefore, we applied the Portmanteau test to test several autocorrelations of r_t that are simultaneously at zero to determine whether the gene is a SVG. The null and alternative hypotheses are:

$$H_0: \rho_1 = \dots = \rho_m = 0, H_A: \exists k \in \{1, \dots, m\}, \rho_k \neq 0$$

To simplify the symbolic representation, we rewrite the subscript of the marginal expression series as $\mathbf{r} = (r_1, \dots, r_t, \dots, r_T)^T$, define the autocovariance of order m as:

$$\gamma_k = Cov(r_t, r_{t-k}) = Cov(r_t, r_{t+k}), \text{ for all } k \geq 0, \text{ and the } j\text{th order autocorrelation (ACF) as } \rho_k = \frac{\gamma_k}{\gamma_0}.$$

If gene g is non-SVG without a spatial pattern in ST data, our purpose is to test the null hypothesis: $H_0: \rho_1 = \dots = \rho_m = 0$.

The test statistic is defined as $Q_m = T \sum_{k=1}^m \hat{\rho}_k^2$ followed by chi-distribution with m degree of freedom, where $\hat{\gamma}_k = \frac{1}{T-k} \sum_{t=1+k}^T (r_t - \bar{r})(r_{t-k} - \bar{r}), k = 0, \dots, T-1$, $\bar{r}$ is the mean of $\mathbf{r}$, and introduce $\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$. The P value for testing the null hypothesis can be

calculated by $p = P(\chi^2(df = m) > Q_m | H_0 \text{ is true})$

Stouffer's method

We combined all four P values into a single P value by Stouffer's method. The Stouffer's

statistic is defined as $z_{stouffer} = \sum_{i=1}^4 \frac{z_i}{\sqrt{4}} \sim N(0,1)$, where $z_i = \Phi^{-1}(1 - p_i)$, $\Phi^{-1}(\cdot)$ is

the inverse of the cumulative distribution function of a standard normal distribution. Hence,

the combined p of four p-values is calculated by $p_c = 1 - \Phi(z)$.

2.2 Semi-pooling process

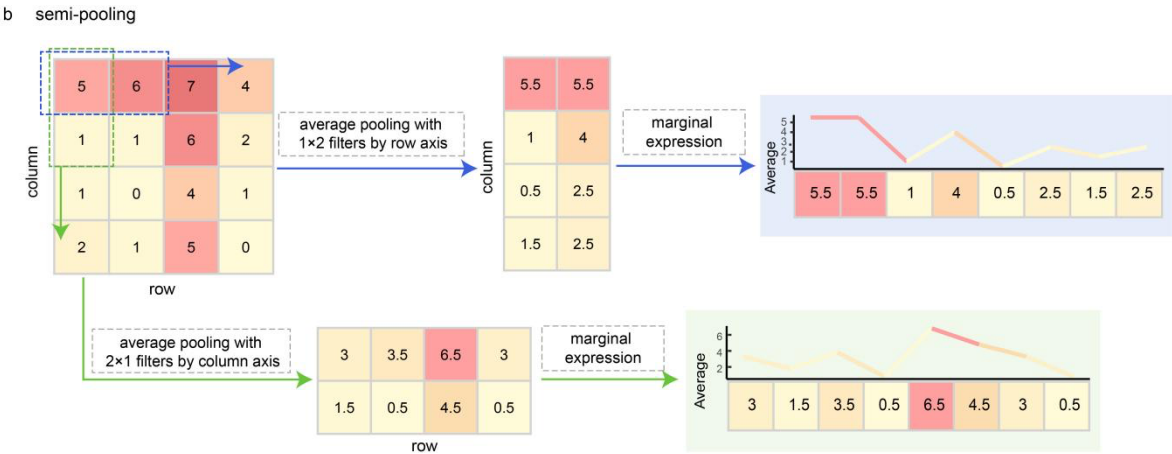


Figure S8 Semi-pooling process

3. Application to colorectal cancer data by 10X Visium



Figure S9 Top 10 SVGs identified by each method. The top 10 genes identified by

HEARTSVG and SPARK-X showed stronger spatial expression patterns compared to SpatialDE and SPARK (Fig.S10-17). SpatialDE's top 10 selected SVGs exhibited minimal spatial patterns, whereas SPARK performed better than SpatialDE to a certain extent.

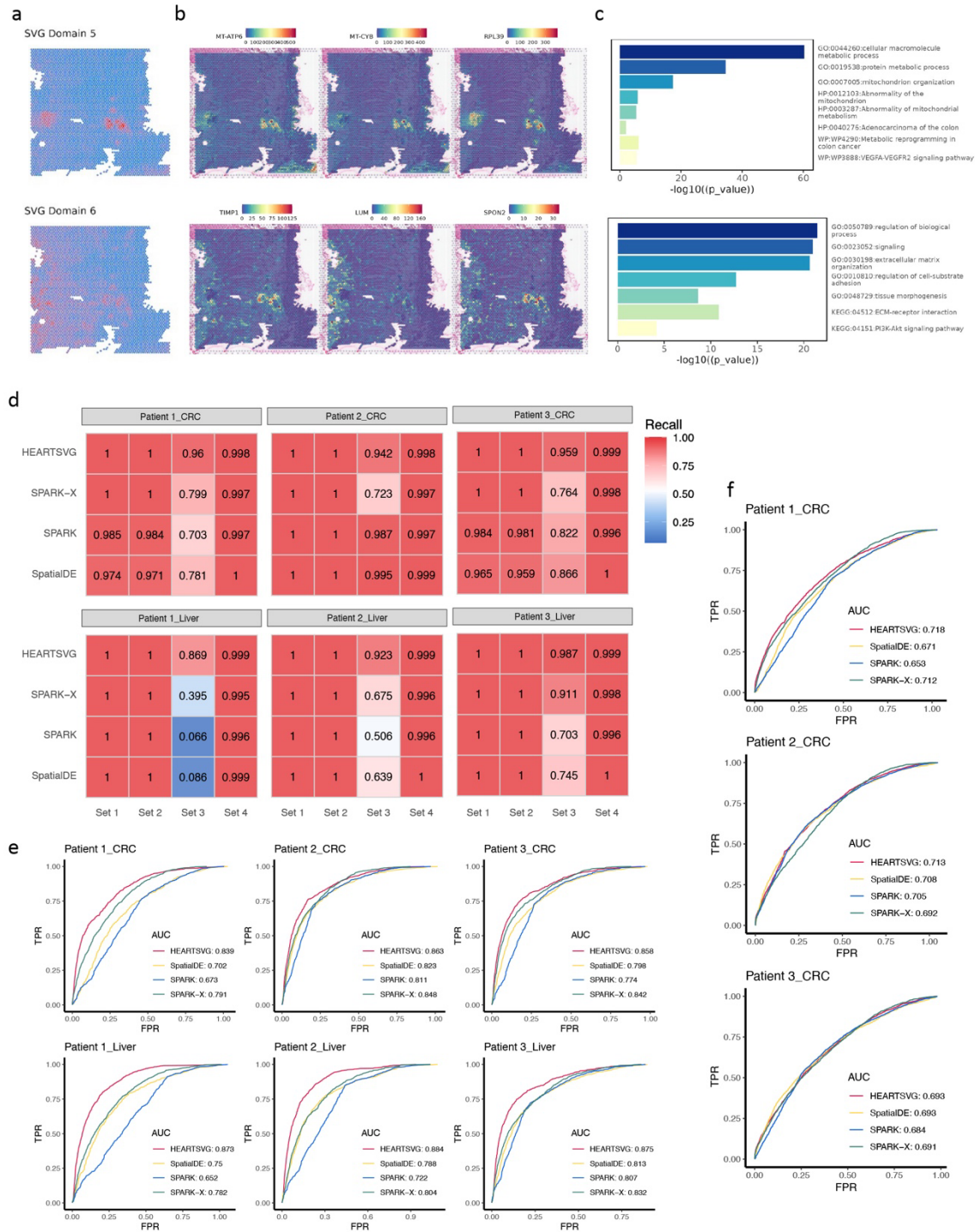


Figure S10 (a) Predicted spatial domain 5,6 based on SVGs. (b) Representative genes of spatial domain 5,6. (c) Enrichment analysis of spatial domain 5,6. (d) The heatmap shows the comparison of recall values among four genesets on three colorectal cancer ST datasets and three corresponding liver metastasis ST datasets. Set 1 represents the true SVGs set,

derived by selecting the top 500 overlaps between HEARTSVG and SPARK-X. Set 4 corresponds to the non-SVGs set, obtained by randomly rearranging the gene expressions within Set 1. Set 2 and Set 3 are generating by introducing noise to the true SVGs set and non-SVGs set, respectively. **(e)** ROC curves were used to assess the true positive rate (TPR) and false positive rate (FPR) of four different methods, using common gene modules of tumor microenvironments as gold standards for true spatially variable genes (SVGs). The figures consisted of six sub-figures representing three colorectal cancer ST datasets and three corresponding liver metastasis ST datasets. Each method was represented by a different colored line, and the area under the ROC curve (AUC) was calculated. **(f)** ROC curves were used to evaluate the TPR and FPR of four different methods, using consensus molecular markers of colorectal cancer subtypes as gold standards for true SVGs. The figures comprised three sub-figures corresponding to three colorectal cancer ST datasets.

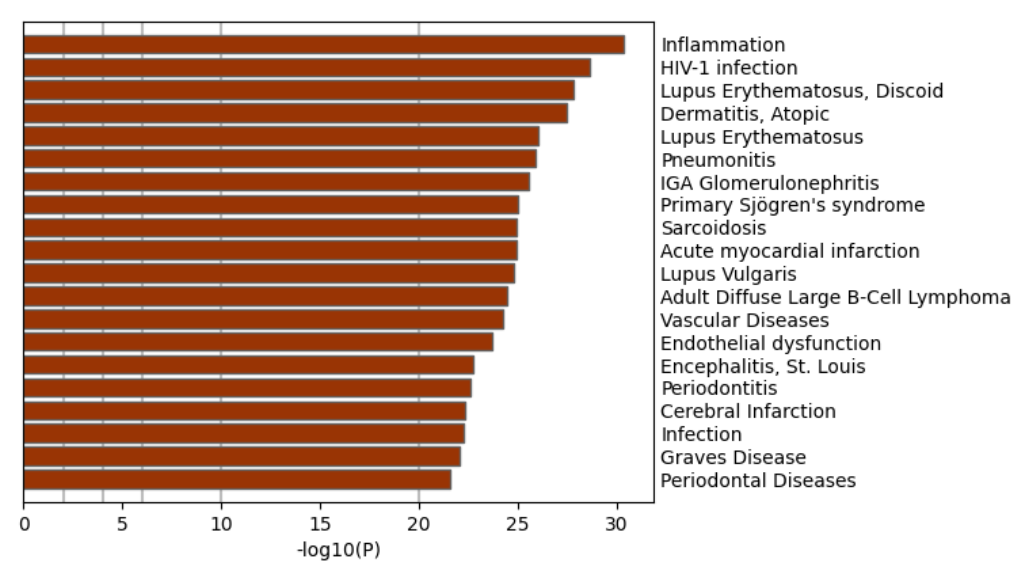


Figure S11 Enrichment results of Spatial domain 1.

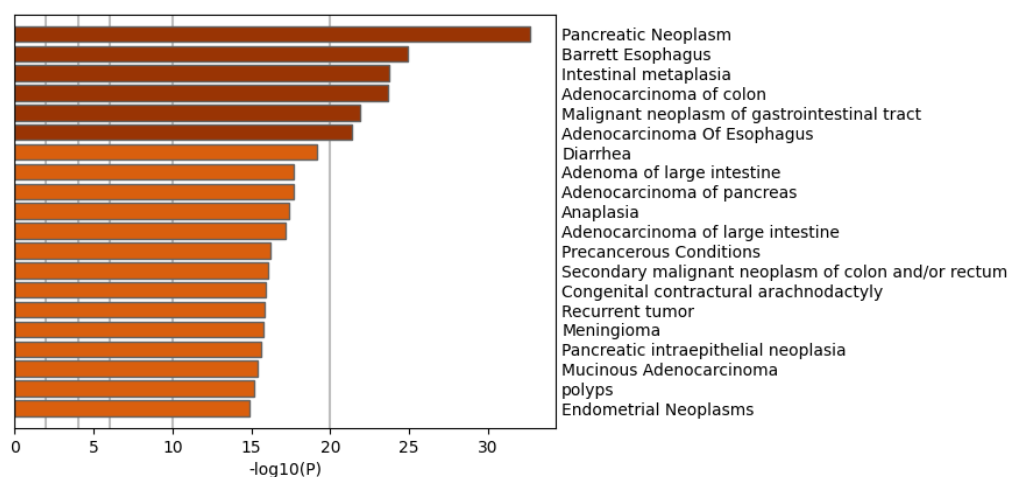


Figure S12 Enrichment results of Spatial domain 2.

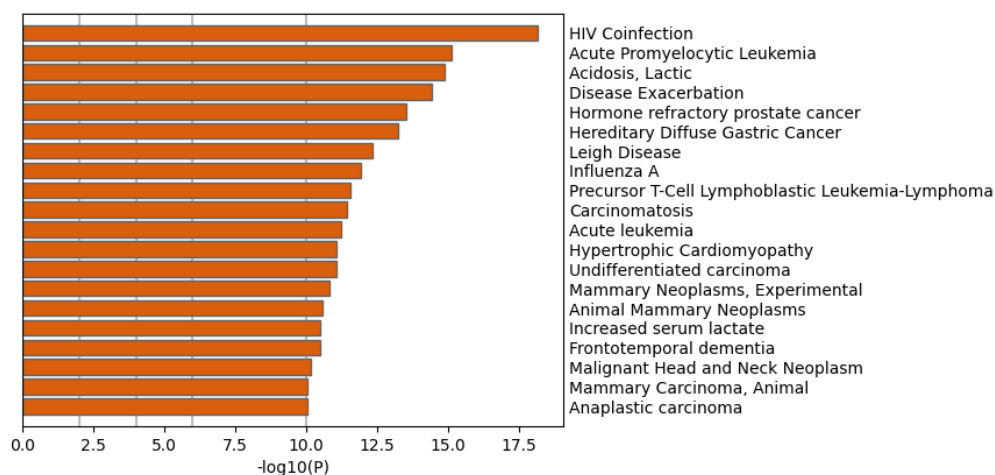


Figure S13 Enrichment results of Spatial domain 3.

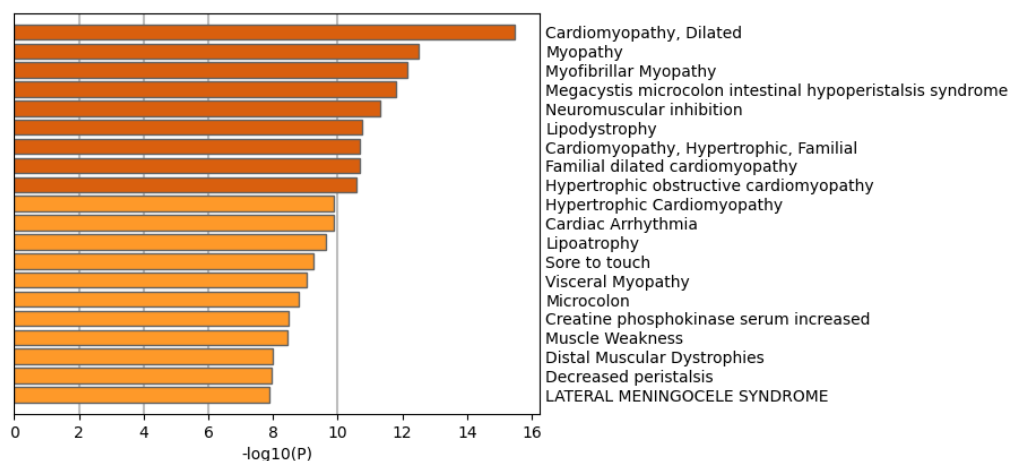


Figure S14 Enrichment results of Spatial domain 4.

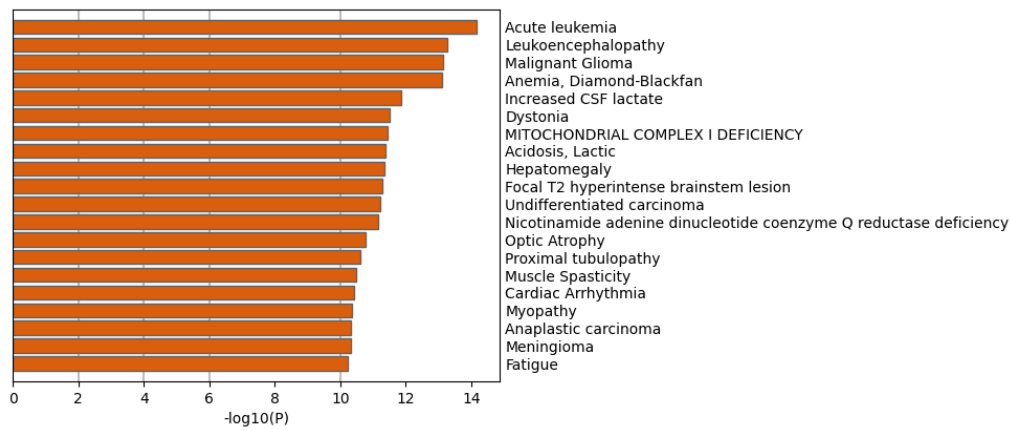


Figure S15 Enrichment results of Spatial domain 5.

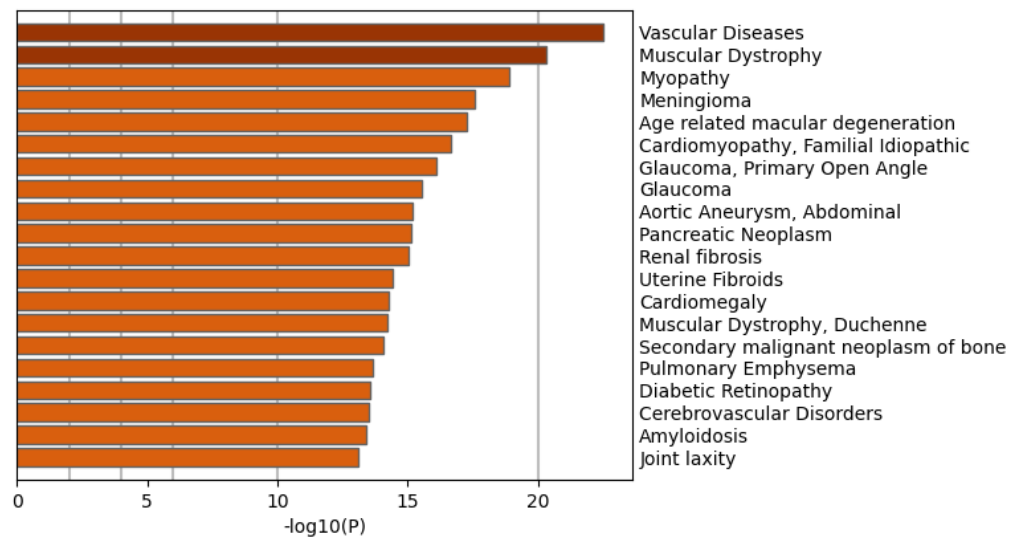


Figure S16. Enrichment results of Spatial domain 6.

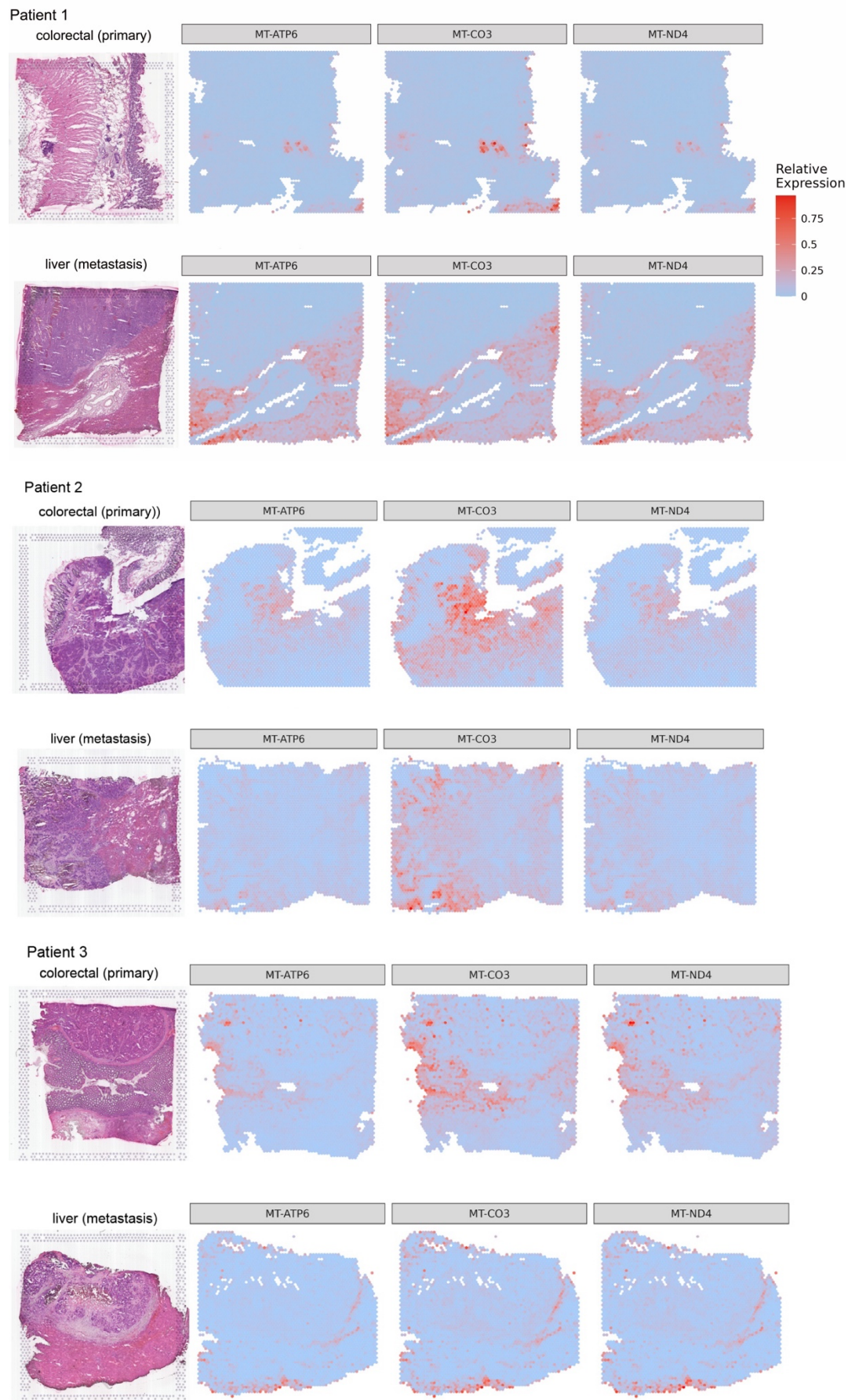


Figure S17 'MT' genes in primary colorectal cancer tissue and liver metastasis cancer tissue.

4. Application to mouse cerebellum data by Slide-seqV2

Table S1 Tissue-specificity enrichment analysis results of each method.

	caudate	cerebellum	cerebral cortex	hippocamp us	rectum	skin	SUM
HEARTSVG	0 (0%)	35 (87.5%)	4 (10%)	1 (2.5%)	0 (0%)	0 (0%)	40
SPARK	5 (5.1%)	42 (42.86%)	7 (7.14%)	4 (4.08%)	3 (3.06%)	37.76(18.3 7%)	98
SPARK-X	1 (2%)	38 (76%)	4 (8%)	1 (2%)	3 (6%)	3 (6%)	50
SpatialDE	0 (0%)	44 (86.27%)	4 (7.84%)	2 (3.92%)	1 (1.96%)	0 (0%)	51

Table S2 Tissue-specificity enrichment analysis results of each method.

gene	method	rank	adjusted p
Car8	HEARTSVG	66	0.000 ***
	SpatialDE	37	0.000 ***
	SPARK	3	0.000 ***
	SPARK-X	585	0.050
Pcp2	HEARTSVG	38	0.000 ***
	SpatialDE	145	0.000 ***
	SPARK	110	0.000 ***
	SPARK-X	804	0.139
Pcp4	HEARTSVG	27	0.000 ***
	SpatialDE	48	0.000 ***
	SPARK	5	0.000 ***
	SPARK-X	2485	1.000

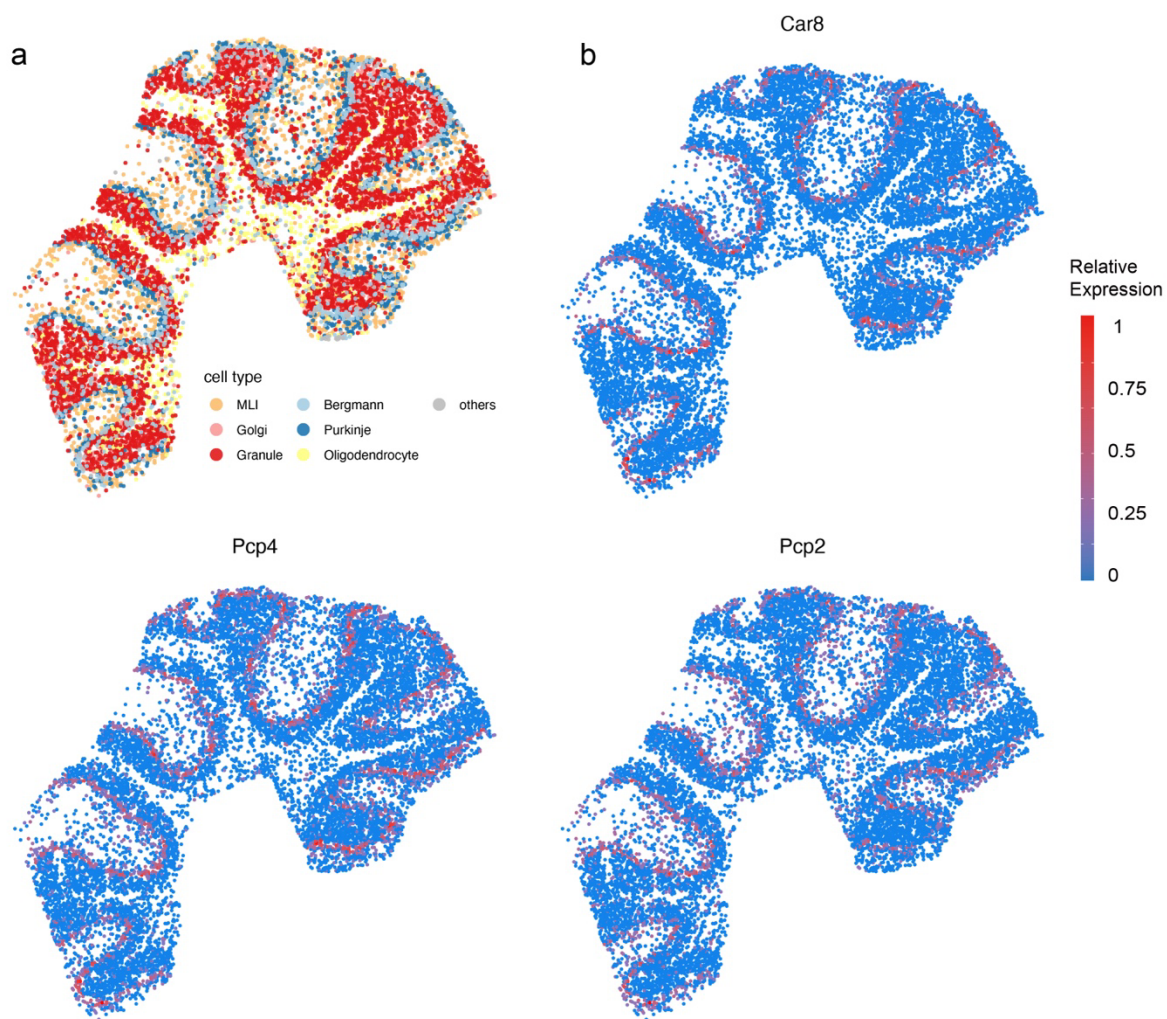


Figure S18 (a) Visualization of unsupervised spatial clustering results. **(b)** Visualizations of marker genes of Purkinje cells in the MERFISH data.

5. Application to mouse preoptic hypothalamus data by MERFISH

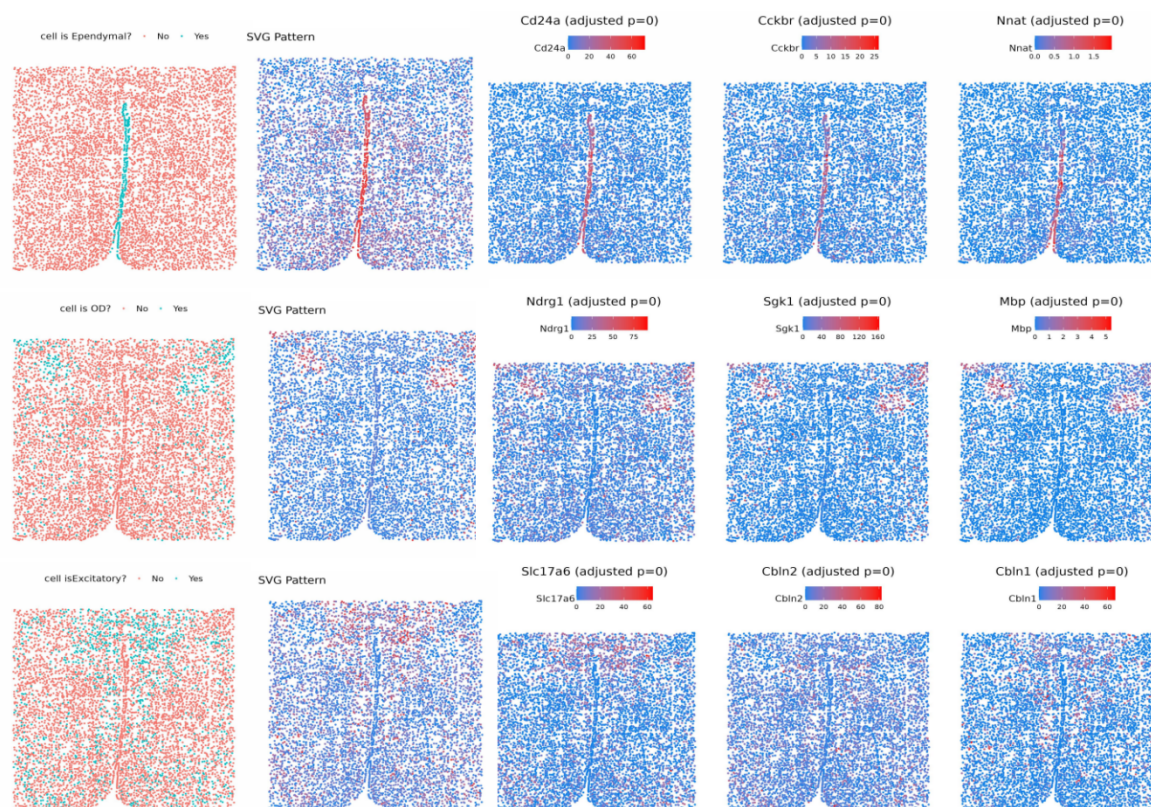
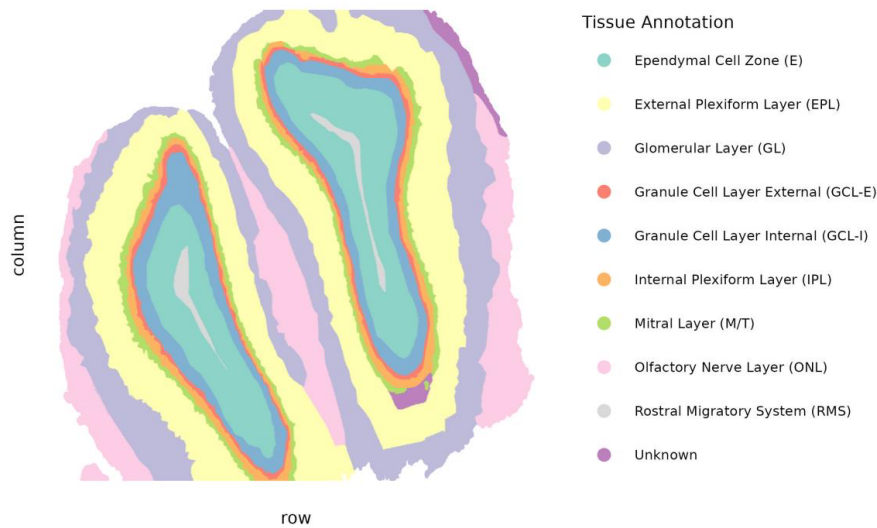


Figure S19 Visualizations of marker genes and cell type of the MERFISH data.

6. Application to mouse olfactory bulb data by HDST

a



b

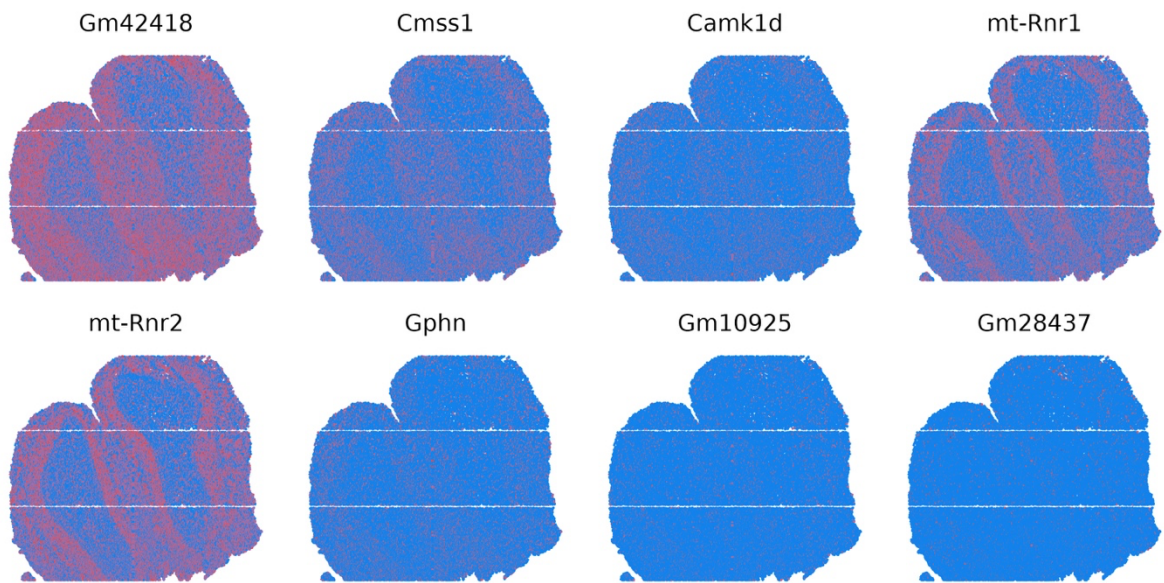


Figure S20 (a) Cell annotations of HDSTdata. (b) Representative SVGs.

7. Application to primary liver cancer data by 10X Visium

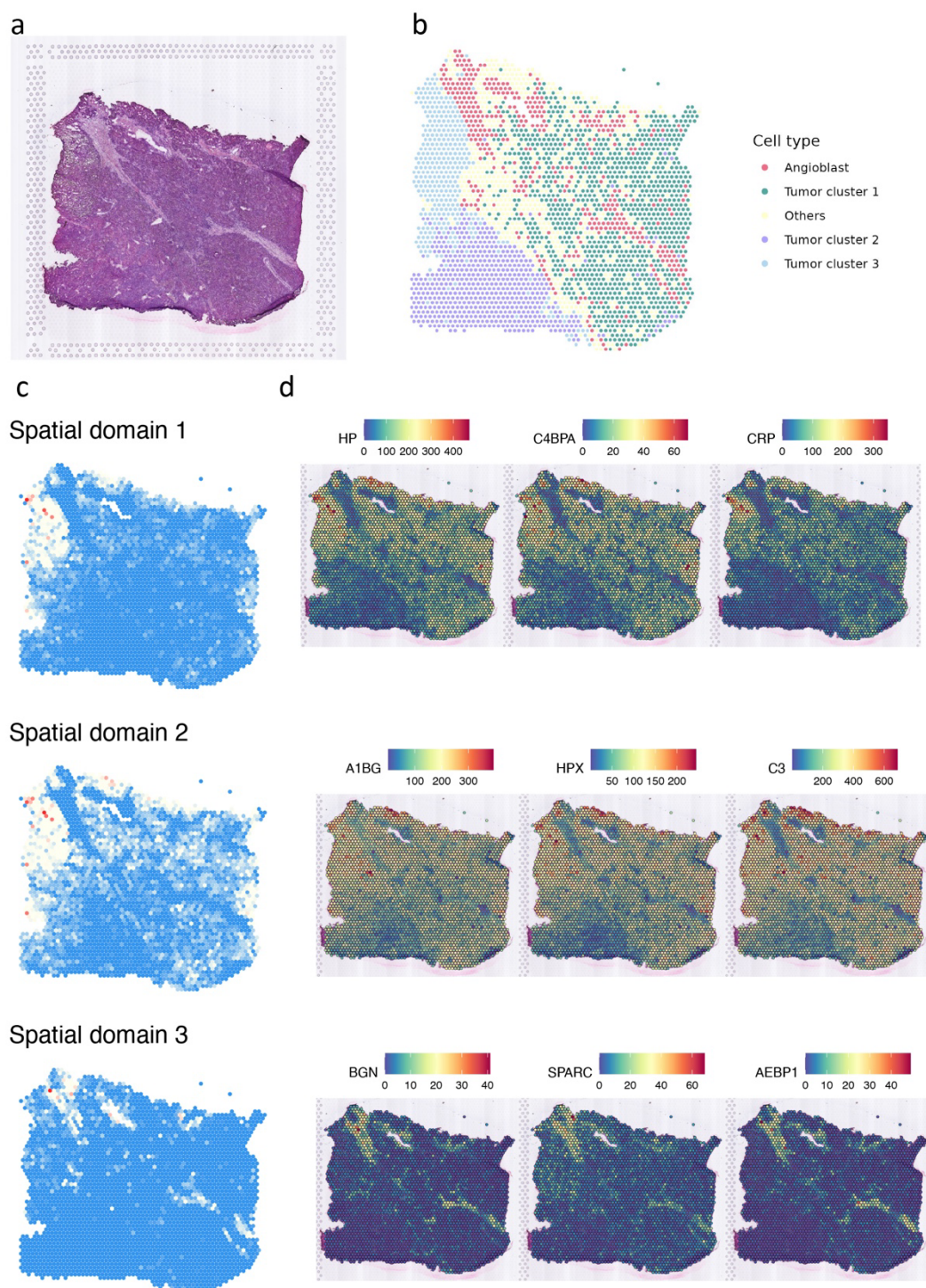


Figure S21 (a) Original hematoxylin and eosin stained (H&E) tissue image. (b) Unsupervised spatial clustering results. (c) SVGs cluster patterns. (d) Representative genes of six SVG clusters.

8. Application to prenatal clear cell cancer brain metastasis data by 10X Visium

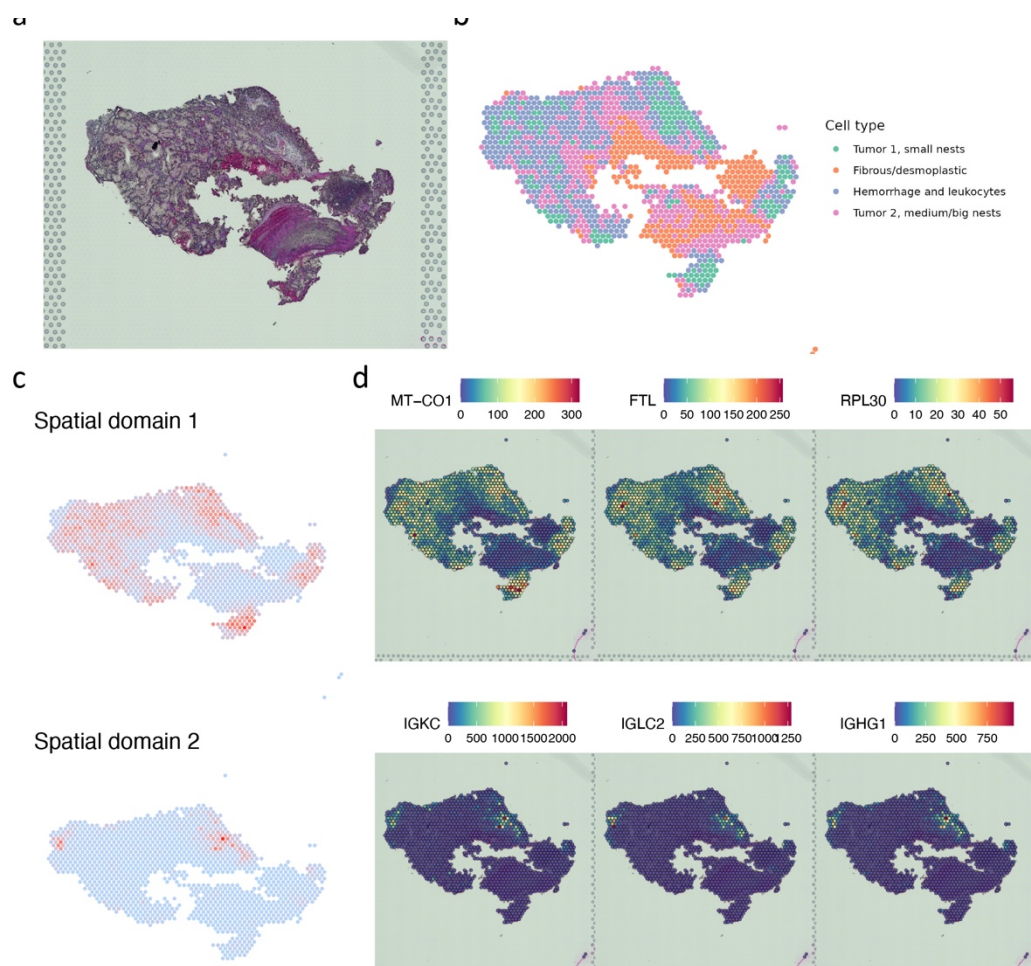


Figure S22 (a) Original hematoxylin and eosin stained (H&E) tissue image. (b) Unsupervised spatial clustering results. (c) SVGs cluster patterns. (d) Representative genes of six SVG clusters.