Identification of Cell-Type-Specific Spatially Variable Genes Accounting for Excess Zeros

Jinge Yu, Xiangyu Luo*

Institute of Statistics and Big Data, Renmin University of China

Abstract

Spatial transcriptomic techniques can profile gene expressions while retaining the spatial information, thus offering unprecedented opportunities to explore the relationship between gene expression and spatial locations. The spatial relationship may vary across cell types, but there is a lack of statistical methods to identify cell-type-specific spatially variable (SV) genes by simultaneously modeling excess zeros and cell-type proportions. We develop a statistical approach CTSV to detect cell-type-specific SV genes. CTSV directly models spatial raw count data and considers zero-inflation as well as overdispersion using a zero-inflated negative binomial distribution. It then incorporates cell-type proportions and spatial effect functions in the zero-inflated negative binomial regression framework. The R package **pscl** (Zeileis et al., 2008) is employed to fit the model. For robustness, a Cauchy combination rule is applied to integrate p-values from mutliple choices of spatial effect functions. Simulation studies show that CTSV not only outperforms the competing methods at the aggregated level but also achieves more power at the cell-type level. By analyzing pancreatic ductal adenocarcinoma spatial transcriptomic data, SV genes identified by CTSV reveal meaningful biological insights at the cell-type level. The R package to implement CTSV is available on GitHub `https://github.com/jingeyu/CTSV`.

Keywords: cell-type-specific, multiple testing, spatial transcriptomics, zero-inflated negative binomial regression

*corresponding author, xiangyuluo@ruc.edu.cn

1 Introduction

The development of spatial transcriptomic techniques has enabled the measurement of gene expression with accompanied spatial context information (Larsson et al., 2021; Zhuang, 2021; Close et al., 2021), providing unprecedented opportunities to investigate the interaction between expression and spatial locations. One crucial challenge in the spatial expression data analysis is to identify genes whose expression levels vary with spatial coordinates in a tissue section, which are termed as spatially variable (SV) genes. In recent years, the task of SV gene detection draws much attention from bioinformaticians, and several statistical methods (Edsgärd et al., 2018; Svensson et al., 2018; Sun et al., 2020; Zhu et al., 2021; Hao et al., 2021; Li et al., 2021) have been proposed to test the dependence of expression on spatial locations. However, the dependence may be confounded by some biological or technical factors, thus resulting in many false positives. In this paper, we aim to mitigate the confounding issues in SV gene identification by accounting for two possible confounding factors—cell-type proportions and excessive zeros.

On the one hand, the commonly used spatial transcriptomics (ST) platforms, including ST based on spatially barcoded microarrays (Ståhl et al., 2016), 10x Genomics Visium (Rao et al., 2020), and Slide-seq (Rodriques et al., 2019), profile gene expression from spots that are regularly organized in a grid in a tissue section. Each spot usually consists of dozens of cells, so the observed expression measurements are at the bulk level rather than at single-cell resolution. Since spots in different tissue regions often have different cell-type proportions (Cable et al., 2021; Elosua-Bayes et al., 2021), the latent cellular compositions can induce expression variations even though the spatial locations have no impact on the expression, thus confounding the SV gene detection. In fact, the confounding issue by cell-type proportions has been also observed in other types of association studies, e.g., the epigenome-wide association studies (Zheng et al., 2018; Luo et al., 2019; Rahmani et al., 2019). On the other hand, unlike traditional bulk RNA-seq or microarray data, the bulk ST expression still suffers from zero-inflation because the expression signals for a large proportion of genes within each spot are too weak to be captured by ST technologies. Figure 1(a) shows a bar plot of spot-wise zero proportions in a real bulk ST dataset (Moncada et al., 2020), and we can observe that more than 80% of spots have at least 70% zeros in the expression. Therefore, it is necessary to account for cell-type proportions and sparsity when modeling bulk ST data.

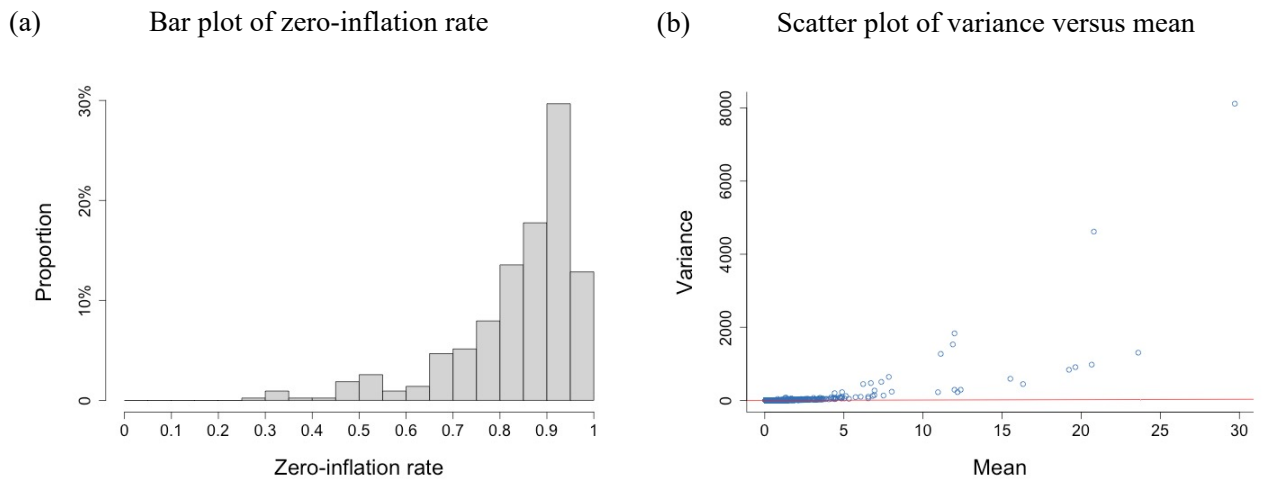

Fig. 1: Zero-inflation and overdispersion in the pancreatic ductal adenocarcinoma (PDAC) ST data. (a) Bar plot of spot-wise zero proportions. (b) Scatter plot of genes' expression variance versus expression mean in PDAC data. Each point corresponds to one gene, and the red solid line represents the line with intercept zero and slope one.

If we do not account for the two factors simultaneously, there has been some statistical works on the identification of SV genes, where frequentist methods carry out multiple hypothesis testings (non-SV in the null and SV in the alternative) and determine the p-value threshold by controlling the false discovery rate (FDR), and Bayesian methods calculate the posterior probability of being SV for each gene using posterior samples and identify SV genes based on estimated Bayesian FDR. Specifically, to our knowledge, trendsceek (Edsgärd et al., 2018) and SpatialDE (Svensson et al., 2018) are the first two statistical methods to achieve that. Trendsceek (Edsgärd et al., 2018) was built upon the marked point process to test whether the joint probability of expressions on two locations relies on their distance, calling it a mark segregation. It then makes use of four types of mark-segregation summary statistics to compute p-values through permutations. As trendsceek models the probability density, it can capture spatial expression changes both from mean and covariance. In contrast, SpatialDE (Svensson et al., 2018) only models the spatial covariance structure using zero mean Gaussian process (Williams and Rasmussen, 2006) and fits spatial expression data via a normal distribution, and then compares the result against a null model without spatial effects to calculate p-values. Recently, Hao et al. (2021) proposes SOMDE using self-organizing map to enhance the computational scalability on large-scale data. However, these methods need to first transform raw expression count data to continuous values, and this may lose power in the downstream analysis (Sun et al., 2017).

SPARK (Sun et al., 2020) is an elegant and powerful statistical method that directly fits spatial raw counts via the Poisson log linear regression model and uses the zero mean Gaussian process to model spatial effects. Hence, it can achieve more power than trendsceek and SpatialDE. It also maintains robustness by considering multiple kernel choices of the Gaussian process and combining multiple p-values through a Cauchy combination rule (Liu et al., 2019). Nevertheless, a simple Poisson distribution cannot account for excess zeros (Figure 1(a)) and overdispersion (Figure 1(b)) in the ST expression data. Recently, BOOST-GP (Li et al., 2021) explicitly models the sparse spatial expression via a zero-inflated negative binomial distribution, where the negative binomial mean is connected to covariates through a log link. Spatial effects are further incorporated via zero mean Gaussian process, and binary indicators are introduced for SV genes. Subsequently, the inference is performed in the Bayesian framework, and the posterior samples of SV gene indicators are used to calculate the posterior inclusion probability. Finally, SV genes are selected based a controlled estimated Bayesian FDR.

Instead of the explicit introduction of zero-inflation in BOOST-GP, Zhu et al. (2021) designs a nonparametric approach SPARK-X that does not need to specify the distribution of sparse spatial expression. SPARK-X extends the scalability of SPARK and further improves its robustness on large scale spatial transcriptomic data. Moreover, as far as we know, currently SPARK-X (Zhu et al., 2021) is the unique SV gene detection method that provides a way to identify cell-type-specific SV genes. Specifically, when applied to Slide-seq v2 data and HDST data, SPARK-X first uses the cell-type proportion estimates from RCTD (Cable et al., 2021) to assign each spot to its major cell type and then detects SV genes for spots of the same labeled cell type. Nevertheless, the assignment procedure ignores the influence of minor cell types in each spot, and thus it is more reasonable to directly utilize the cell-type proportion estimates to identify cell-type-specific SV genes.

In this paper, we develop a simple statistical approach CTSV to identify Cell-Type-specific SV genes accounting for excess zeros. CTSV directly fits the sparse expression raw counts using a zero-inflated negative binomial distribution, models the mean as a weighted average of cell-type-specific spatial expression profiles with weights being the cell-type proportions, and for each cell type connects the spatial expression profile to a function of spatial coordinates. By combining these equations in CTSV, the identification of cell-type-specific SV genes is equivalent to testing whether the function of spatial coordinates is zero for each cell type in

a zero-inflated negative binomial regression model. Specifically, since there has been several mature bulk ST deconvolution methods (Cable et al., 2021; Elosua-Bayes et al., 2021; Dong and Yuan, 2021), we treat the estimated cell-type proportions as fixed covariates in CTSV. We further model unknown functions to be linear, focal, and periodic, respectively, and combine the p-values from the multiple choices to achieve the robustness to unavailable spatial pattern like in SPARK (Sun et al., 2020). Through simulation studies, CTSV can achieve more power than SPARK-X in detecting cell-type-specific SV genes and also outperforms other methods at the aggregated level. The real data analysis to PDAC ST data also shows the practical utility of CTSV.

2 Method

2.1 The Proposed Approach CTSV

Suppose there are G genes, n spots, and K cell types in the tissue section. Assume that $\mathbf{Y} = \{Y_{gi} : 1 \leq g \leq G, 1 \leq i \leq n\}$ is the bulk ST data matrix, where Y_{gi} is the observed raw count of gene g in spot i. Let $\mathbf{S} = \{(s_{i1}, s_{i2}) : 1 \leq i \leq n\}$ represent the set of coordinates of spots' centers, and $\boldsymbol{s}_i = (s_{i1}, s_{i2})$ is the two dimensional coordinate of spot i's center. To account for the count nature and overdispersion of ST data, we consider the negative binomial distribution $\text{NB}(c_i\lambda_{gi}, \psi_g)$ with mean $c_i\lambda_{gi}$ and shape parameter ψ_g for gene g in spot i, and its probability mass function is $f(x|c_i\lambda_{gi}, \psi_g) = \frac{\Gamma(x+\psi_g)}{x!\cdot\Gamma(\psi_g)} \frac{(c_i\lambda_{gi})^x \cdot \psi_g^{\psi_g}}{(c_i\lambda_{gi}+\psi_g)^{x+\psi_g}}$ for any non-negative integer x. In this way, the variance equals $c_i\lambda_{gi} + (c_i\lambda_{gi})^2/\psi_g$ and thus is larger than the mean $c_i\lambda_{gi}$. The scalar c_i is a size factor to account for different library sizes of spots, and it is computed to be the ratio of spot i's library size to the median library size across spots, i.e., $c_i = \frac{\sum_{g=1}^{G} Y_{gi}}{\text{median}_{1\leq j\leq n} \sum_{g=1}^{G} Y_{gj}}$.

In addition to overdispersion, bulk ST data also suffer from zero-inflation—the observed zero proportion is much larger than the expected zero proportion of a negative binomial distribution. Typically there are two kinds of zeros in the data. One is called "biological zeros" resulting from genes that do not express, and the other one is "technical zeros," which means that some genes have relatively low expressions and thus are not captured. Taking both overdispersion and zero-inflation into consideration, we model the count data Y_{gi} by a zero-inflated negative binomial distribution,

$$Y_{gi} \sim \pi_g\delta_0 + (1 - \pi_g)\text{NB}(c_i\lambda_{gi}, \psi_g), \tag{1}$$

where π_g denotes the probability of being a technical zero for gene g in the spots and δ_0 is a Dirac measure with point mass at zero.

As one spot may consist of dozens of heterogeneous cells, we model the log scale of λ_{gi} as a mix of cell-type-specific expression levels of gene g in spot i,

$$\log \lambda_{gi} = \sum_{k=1}^{K} \mu_{gki} w_{ik}. \tag{2}$$

w_{ik} is the cell-type k proportion in spot i, and μ_{gki} represents the expression level of gene g for cell type k in spot i. μ_{gki} depends on the spot i through its location $\mathbf{s}_i$, and the relationship is modeled as follows using a similar formulation from Luo et al. (2019).

$$\mu_{gki} = \eta_{gk} + \beta_{gk1} h_1(s_{i1}) + \beta_{gk2} h_2(s_{i2}), \tag{3}$$

where η_{gk} is the cell-type-k baseline expression level of gene g, the two functions $h_1(\cdot)$ and $h_2(\cdot)$ describe the spatial effects on the mean η_{gk}, and the coefficients β_{gk1} and β_{gk2} are of our interest that can reflect whether the location $\mathbf{s}_i$ affects the expression of gene g in cell type k. Subsequently, by combining Equations (1)-(3), we arrive at the proposed approach CTSV (Cell-Type-specific Spatially Variable gene detection),

$$\begin{aligned} Y_{gi} &\sim \pi_g \delta_0 + (1-\pi_g)\mathrm{NB}(c_i \lambda_{gi}, \psi_g), \\ \log \lambda_{gi} &= \sum_{k=1}^{K} \mu_{gki} w_{ik}, \\ \mu_{gki} &= \eta_{gk} + \beta_{gk1} h_1(s_{i1}) + \beta_{gk2} h_2(s_{i2}). \end{aligned}$$

If we integrate the last two equations, CTSV is equivalent to

$$\begin{aligned} Y_{gi} &\sim \pi_g \delta_0 + (1-\pi_g)\mathrm{NB}(c_i \lambda_{gi}, \psi_g), \\ \log \lambda_{gi} &= \sum_{k=1}^{K} \eta_{gk} \cdot w_{ik} + \sum_{k=1}^{K} \beta_{gk1} \cdot h_1(s_{i1}) w_{ik} + \sum_{k=1}^{K} \beta_{gk2} \cdot h_2(s_{i2}) w_{ik}. \end{aligned} \tag{4}$$

Our next goal is to conduct statistical inference for the coefficients β_{gk1} and β_{gk2} to test whether they are zero or not for each gene. Specifically, if at least one of these null hypotheses $H_0: \ \beta_{gk1} = 0$ and $H_0: \ \beta_{gk2} = 0$ is rejected, then we believe that gene g is SV in cell type k.

2.2 Statistical Inference

2.2.1 When functions h_1 and h_2 are known

In Equation (4), if we know the cellular compositions $\{w_{ik} : k = 1, \ldots, K\}$ for each spot i as well as the functions h_1 and h_2, then we can treat them as covariates and thus the inference for CTSV reduces to the inference for a zero-inflated negative binomial regression model (Preisser et al., 2016), which can be easily conducted by the R package **pscl** (Zeileis et al., 2008). However, the cellular compositions of each spot are often unavailable. Fortunately, there has been several deconvolution methods designed for bulk ST data recently, such as RCTD (Cable et al., 2021), SPOTlight (Elosua-Bayes et al., 2021), and SpatialDWLS (Dong and Yuan, 2021). Subsequently, we treat the estimates for $\{w_{ik} : k = 1, \ldots, K\}$ as fixed covariates and plug them in Equation (4).

Next, based on the R package **pscl** (Zeileis et al., 2008), we can obtain the p-value $p_{gk\ell}$ for the hypothesis $H_0 : \beta_{gk\ell} = 0$ vs $H_1 : \beta_{gk\ell} \neq 0$ for gene g in cell type k along the ℓth coordinate ($\ell = 1, 2$). Notice that as the inference is carried out for each gene independently, the procedure is highly parallellable. All the p-values can be organized into a p-value matrix $\{p_{gk\ell}\}$ with dimension $G \times 2K$, where the kth ($1 \leq k \leq K$) column corresponds to the p-value vector in cell type k for the s_1 coordinate and the $(K + k)$th ($1 \leq k \leq K$) column to the p-value vector in cell type k for the s_2 coordinate. To control the false discovery rate (FDR) in the multiple hypothesis testings, we convert the p-value matrix to the q-value matrix $\{q_{gk\ell}\}_{G\times 2K}$ using the R package **qvalue** (Storey and Tibshirani, 2003; Storey et al., 2020). In this way, a q-value threshold α controls the false discovery rate to be not larger than α.

Specifically, for each g-th row in the q-value matrix, if there is at least one q-value in this row ($q_{gk\ell} : 1 \leq k \leq K,\ \ell = 1, 2$) less than α, we call the corresponding gene g SV at the aggregated level. For each cell type k, if there is at least one q-value in ($q_{gk\ell} : \ell = 1, 2$) less than α, we then identify the gene g to be cell-type-k-specific SV.

2.2.2 When functions h_1 and h_2 are unknown

In practice, we often do not know what the type of underlying spatial patterns is in the tissue section for each gene. To deal with possible model misspecification and make the CTSV method more robust, we follow the idea from Sun et al. (2020) to choose three types of functions for h_1 and h_2, which can reflect the linear, focal, and periodic spatial expression patterns. Specifically,

suppose that $\mathbf{s}_1$ and $\mathbf{s}_2$ are first transformed to have mean zero and standard deviation one. We choose linear functions as $h_1(s_{i1}) = s_{i1}$ and $h_2(s_{i2}) = s_{i2}$, squared exponential functions $h_{gk1}(s_{i1}) = \exp(-\frac{s_{i1}^2}{2\sigma_1^2})$ and $h_{gk2}(s_{i2}) = \exp(-\frac{s_{i2}^2}{2\sigma_2^2})$, and periodic functions $h_1(s_i) = \cos\left(\frac{2\pi s_{i1}}{\phi_1}\right)$ and $h_2(s_i) = \cos\left(\frac{2\pi s_{i2}}{\phi_2}\right)$. Moreover, for the squared exponential functions, we choose two sets of scale length parameters by (i) letting σ_1 and σ_2 be the 40% quantile of the absolute values of the transformed s_{i1} and s_{i2}, respectively, denoted by $\sigma_1 = Q_{40\%}(|\mathbf{s}_1|)$, $\sigma_2 = Q_{40\%}(|\mathbf{s}_2|)$; and (ii) letting $\sigma_1 = Q_{60\%}(|\mathbf{s}_1|)$, $\sigma_2 = Q_{60\%}(|\mathbf{s}_2|)$. Similarly, for periodic functions, we set (i) $\phi_1 = Q_{40\%}(|\mathbf{s}_1|)$, $\phi_2 = Q_{40\%}(|\mathbf{s}_2|)$ and (ii) $\phi_1 = Q_{60\%}(|\mathbf{s}_1|)$, $\phi_2 = Q_{60\%}(|\mathbf{s}_2|)$. Hence, for each gene g in cell type k along ℓth coordinate, we obtain five p-values.

Accordingly, for gene g in cell type k along ℓth coordinate, we combine the five p-values $(p_{gk\ell}^{(i)} : 1 \le i \le 5)$ following the Cauchy combination rule ACAT (Liu et al., 2019). We first convert each of the five p-values into a Cauchy statistic $T_{gk\ell}^{(i)} = \tan[\pi(0.5 - p_{gk\ell}^{(i)})]$, then take an average of them $T_{gk\ell} = \frac{1}{5}\sum_{i=1}^{5} T_{gk\ell}^{(i)}$, and transform the average into a single p-value $p_{gk\ell} = \mathbb{P}(C \ge T_{gk\ell})$, where C follows the standard Cauchy distribution (Liu et al., 2019; Pillai and Meng, 2016). In this way, we convert five p-value matrices to one p-value matrix $(p_{gk\ell})_{G\times 2K}$, and then the inference is based on the FDR control as discussed in the last subsection.

3 Simulation

In this section, we compared the performance of our method with several state-of-the-art SV gene detection methods. We generated the spatial transcriptomic raw count data following Equation (4), where related parameters are set as follows. Suppose there are $G = 10,000$ genes, $n = 600$ spots, and $K = 6$ cell types. The cell-type-k baseline expression profile $\boldsymbol{\eta}_k$ was generated from normal distributions. Specifically, we first independently simulated η_{g1} from $\mathrm{N}(2, 0.2^2)$ for $g = 1, \ldots, G$ in cell type 1 and then randomly sampled 300 differentially expressed (DE) genes for each cell type k $(2 \le k \le K)$. Next, on the cell-type-k DE genes $(2 \le k \le K)$, we sampled η_{gk} from $\mathrm{N}(\theta_k, \tau_k^2)$ independently, where $(\theta_2, \tau_2) = (3, 0.2)$, $(\theta_3, \tau_3) = (2, 0.2)$, $(\theta_4, \tau_4) = (4, 0.2)$, $(\theta_5, \tau_5) = (3, 0.2)$, $(\theta_6, \tau_6) = (4, 0.2)$. For expressions on the remaining genes, we set $\eta_{gk} = \eta_{g1}$.

Moreover, we partitioned the spot region into four regions as displayed in Figure 2(a) and then sampled cell-type proportions $\boldsymbol{w}_i$ of spot i from Dirichlet distributions. Cell-type proportions of spots in regions from 1 to 4 were independently sampled from $\mathrm{Dir}(1, 1, 1, 1, 1, 1)$,

$\mathrm{Dir}(1,3,5,7,9,11)$, $\mathrm{Dir}(16,14,12,10,8,6)$, and $\mathrm{Dir}(1,4,4,4,4,1)$, respectively. For coefficients β_{gk}, we set 200 SV genes in each cell type, and there were 700 SV genes at the aggregated level. Figure 2(b) shows the SV gene distribution patterns in each cell type. We further consider the following three simulation settings to specify the spatial effects h_1 and h_2.

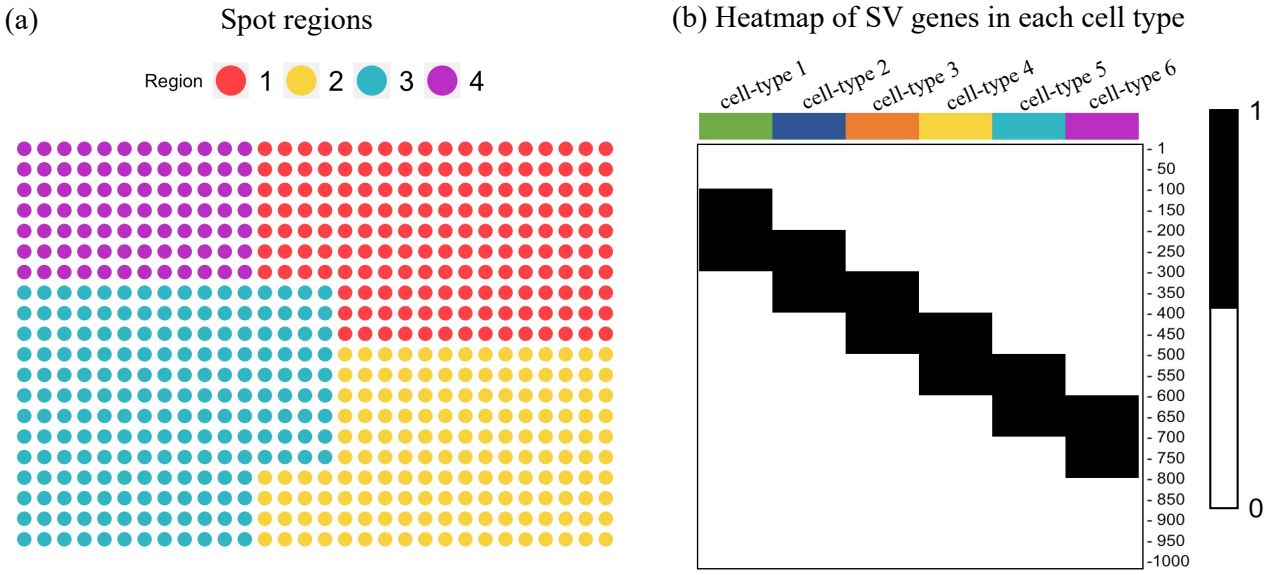

Fig. 2: Spot regions and the heatmap of cell-type-specific SV gene pattern. (a) Four spot regions with different colors. (b) Heatmap of the SV gene pattern. If one gene in a cell type is SV, then it is colored by black. Only the first 1,000 genes are shown for a good visualization because all the remaining genes are not SV.

(1) For the linear spatial pattern as shown in Figure 3(a), we chose $h_1(s_{i1}) = s_{i1}$ and $h_2(s_{i2}) = s_{i2}$. For SV genes, we set $\beta_{gk1} = 1.8$ and $\beta_{gk2} = 0.8$ for each cell type. For non SV genes, $\beta_{gk\ell}$ was set to be zero.

(2) For the focal spatial pattern as shown in Figure 3(b), we set $h_1(s_{i1}) = \exp(-\frac{s_{i1}^2}{2})$ and $h_2(s_{i2}) = \exp(-\frac{s_{i2}^2}{2})$. For SV genes in each cell type, we set $\beta_{gk1} = 3$ and $\beta_{gk2} = 1$. For non SV genes, $\beta_{gk\ell}$ was set to be zero.

(3) For the periodic spatial pattern as shown in Figure 3(c), we have $h_1(s_{i1}) = \cos(2\pi s_{i1})$, $h_2(s_{i2}) = \cos(2\pi s_{i2})$. For SV genes in each cell type, we set $\beta_{gk1} = 2.5$ and $\beta_{gk2} = 1$. For non SV genes, $\beta_{gk\ell}$ was set to be zero.

After obtaining $\boldsymbol{\eta}_k$, $\boldsymbol{w}_i$, $h_1(s_{i1})$, $h_2(s_{i2})$, and $\beta_{gk\ell}$, we can calculate $\log \lambda_{gi}$ and then sample Y_{gi} from $\mathrm{NB}(c_i\lambda_{gi}, \psi_g)$, where the shape parameter is $\psi_g = 100$ and $c_i = 1$. Considering ST data have a large proportion of zeros, we set π_g $(g = 1, \ldots, G)$ to be 0.6 in each spatial pattern.

Therefore, for each gene, the count data was set to be zero with a dropout probability 0.6. In other words, the zero proportions of the simulated data were around 60%. Subsequently, we applied the proposed method CTSV to the three types of simulated ST data and compared the performance with trendsceek (Edsgärd et al., 2018), SpatialDE (Svensson et al., 2018), SPARK (Sun et al., 2020), SPARK-X (Zhu et al., 2021), BOOST-GP (Li et al., 2021), and SOMDE (Hao et al., 2021).

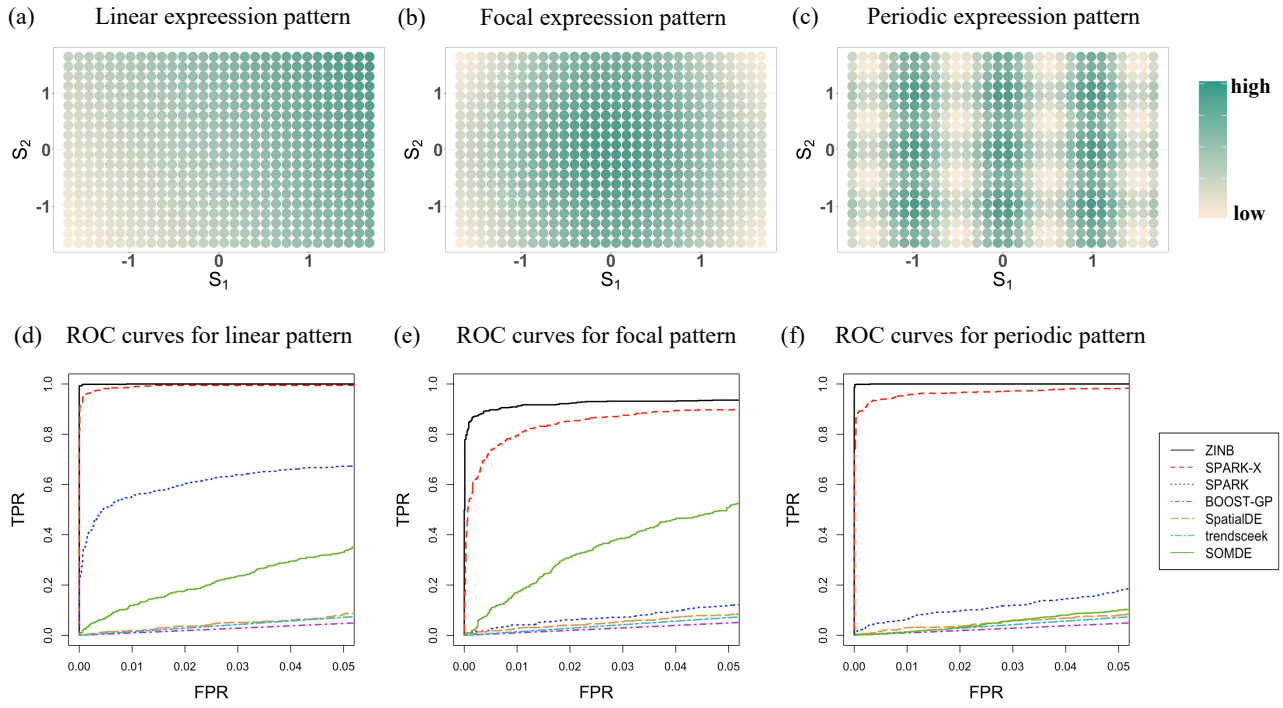

Fig. 3: SV genes' spatial expressions in (a) linear pattern, (b) focal pattern, and (c) periodic pattern, where the coordinates are scaled to have mean zero and standard deviation one. (d-f) The ROC curves with FPR controlled to be less than 0.05 for CTSV, SPARK-X, SPARK, BOOST-GP, SpatialDE, trendsceek, and SOMDE in the three spatial expression patterns.

When implementing CTSV, we considered the estimate error for the cell-type proportions and sampled $\widehat{\boldsymbol{w}}_i$ from $\text{Dir}(\alpha_0 \boldsymbol{w}_i)$ with $\alpha_0 = 100$. In addition, if NA (Not Available) is returned by the function *zeroinfl* in R package **pscl** (Zeileis et al., 2008), the corresponding p-value is recorded as one. In the argument of function *zeroinfl*, some commonly used optimization methods can be used, such as BFGS, conjugate gradient (CG), or Nelder-Mead, and we applied CG algorithm for its stability during the optimization procedure.

The receiver operating characteristic (ROC) curves for identifying SV genes at the aggregated level in the three simulation settings were reported in Figure 3(d)-(f), respectively, where the false positive rate (FPR) is controlled to be less than 0.05 for a good visualization of the

Table 1: True positive rate (TPR) and the number of false positives (FP) for different methods

	Spatial pattern	CTSV	SPARK-X	SPARK	BOOST-GP	SpatialDE	SOMDE	tendsceek
TPR	Linear	0.999	0.907	0.178	0.001	0	0	0
	Focal	0.871	0.293	0	0.001	0	0	0
	Periodic	0.999	0.819	0	0.003	0	0	0
FP	Linear	33	1	0	5	0	0	0
	Focal	19	3	0	5	0	0	0
	Periodic	21	3	0	4	0	0	0

performance comparison. The partial ROC curves indicate that CTSV uniformly outperformed other methods in SV gene detection at the aggregated level. In each setting, the performance of CTSV was followed by SPARK-X, which also performs well due to its nonparametric nature. SPARK ranked the third for the linear and periodic settings, while SOMDE ranked the third in the focal spatial pattern. SpatialDE, trendsceek, and BOOST-GP fail to achieve enough power in all the three simulation settings. Note that trendsceek has four types of statistics, and we only showed the best one. When controlling the FDR less than 0.01 for each method (i.e., the q-value threshold is 0.01), Table 1 demonstrates the true positive rates (TPR) and the number of false positives (FP) in the three spatial expression patterns for all the methods. CTSV and SPARK-X gave much higher TPR than other methods, while the FP of CTSV was slightly larger than SPARK-X. We also observed that trendsceek, SpatialDE, and SOMDE cannot identify any SV gene with FDR less than 0.01. Therefore, at the aggregated level, CTSV can provide a high power with controlled FP and FDR owing to its ability to handle excess zeros and account for cell-type proportions.

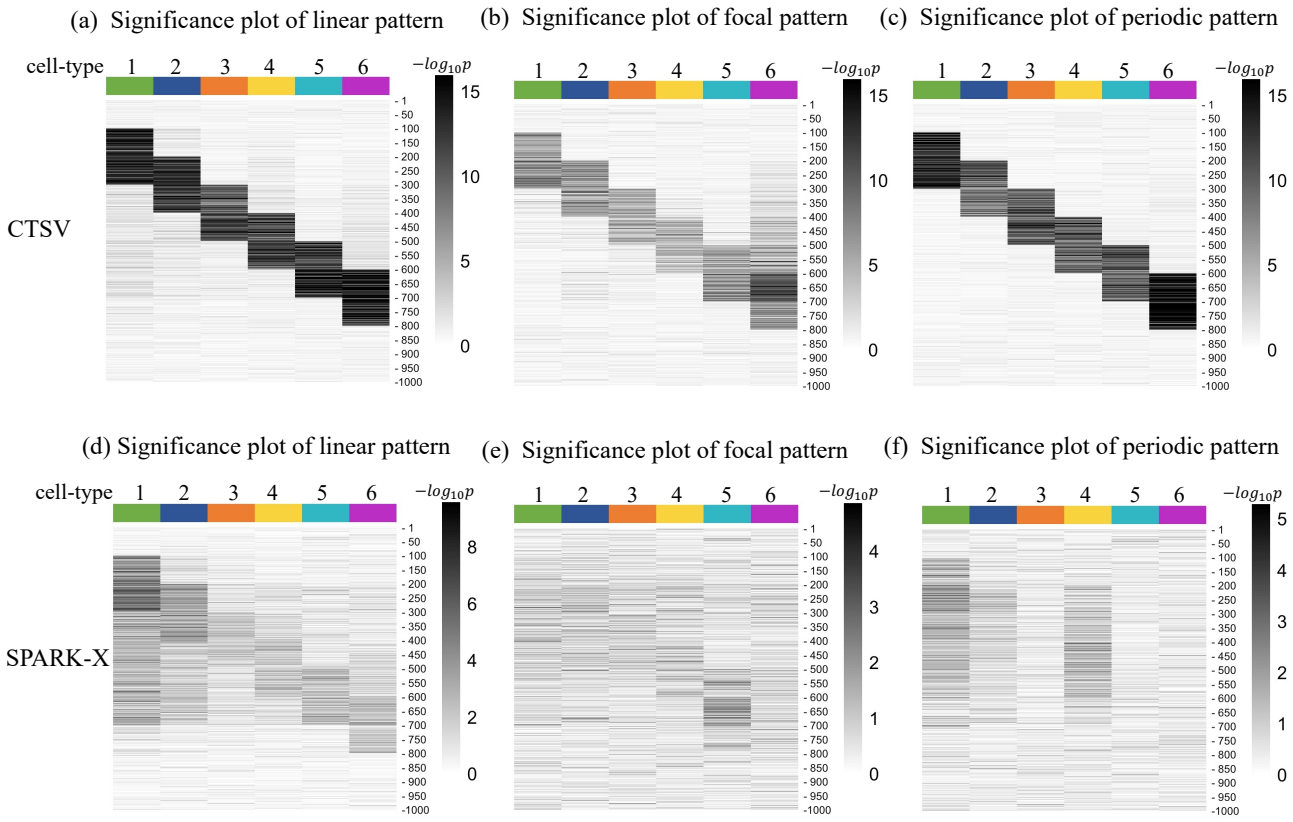

Fig. 4: (a-c) Significance plots of CTSV and (d-f) significance plots of SPARK-X in the three spatial expression patterns for the first 1,000 genes. Values in the heatmaps are $-\log_{10} p$ of the corresponding gene in each cell type. The darker the color, the more likely the corresponding gene is to be SV in that cell type.

Table 2: Cell-type-specific TPR and FP of CTSV and SPARK-X

Spatial pattern		Linear		Focal		Periodic	
Methods		CTSV	SPARK-X	CTSV	SPARK-X	CTSV	SPARK-X
TPR	cell-type 1	1	0.375	0.800	0	1	0
	cell-type 2	0.995	0.095	0.785	0	0.940	0
	cell-type 3	0.980	0	0.605	0	0.970	0
	cell-type 4	0.975	0	0.515	0	0.980	0
	cell-type 5	0.995	0	0.780	0	0.970	0
	cell-type 6	0.995	0	0.905	0	0.995	0
FP	cell-type 1	35	33	9	0	1	0
	cell-type 2	22	4	9	0	1	0
	cell-type 3	10	0	5	0	7	0
	cell-type 4	11	0	8	0	6	0
	cell-type 5	3	0	9	0	3	0
	cell-type 6	21	0	119	0	5	0

Regarding the detection of cell-type-specific SV genes, as SPARK-X is currently the only method that can achieve the function, we compared CTSV and SPARK-X. In SPARK-X (Zhu et al., 2021), if one spot was dominated by a cell type, which has the maximal proportion in that spot, SPARK-X assigned the spot to the cell type. Subsequently, SPARK-X performed the detection task on spots with the same cell type. Figure 4 displays the heatmaps of $-\log_{10}(p_{gk})$ $(g = 1, \ldots, 1,000)$ of CTSV and SPARK-X, where p_{gk} is the p-value of gene g in cell-type k for SPARK-X, and $p_{gk} = \min(p_{gk1}, p_{gk2})$ for CTSV. The darker the color, the more significant that the corresponding gene is SV in that cell type. Compared with the underlying truth (Figure 2(b)), CTSV obtained more accurate results in identifying cell-type-specific SV genes than SPARK-X. Table 2 indicates that when FDR is controlled to be less than 0.01, CTSV yielded higher power than SPARK-X for all the cell types in the three simulation settings, but CTSV did not perform very well in the focal spatial expression pattern. The results showed that CTSV is good at identifying cell-type-specific SV genes by directly modeling cell-type proportions rather than transforming them to one-hot code like in SPARK-X, which may lose some information.

4 Real data analysis

We applied CTSV to pancreatic ductal adenocarcinoma (PDAC) ST data (Moncada et al., 2020), which can be downloaded from Gene Expression Omnibus (Edgar et al., 2002) with accession code GSE111672, and our analysis focuses on the ST1 data from PDAC patient A. As there are associated scRNA-seq data with 18 cell types for patient A, we employed the deconvolution approach SPOTlight (Elosua-Bayes et al., 2021) to obtain cell-type proportion estimates $\widehat{\boldsymbol{w}}_i$ of each spot. We then merged cancer clones A and B into one cell type denoted by "cancer cell," and combined macrophages A and B to one cell type named "macrophages." To alleviate the effects of rare cell types, we calculated the 80th percentile of proportions across spots for each cell type and removed cell types whose 80th percentile is less than 0.1. After the procedure, six cell types—antigen presenting ductal cells, centroacinar ductal cells, high/hypoxic ductal cells, terminal ductal cells, cancer cells, and macrophages—were remained for downstream analysis, and their proportions were adjusted such that they are positive and summed to be one.

Subsequently, we filtered out genes that are expressed in less than 20 spots and kept all

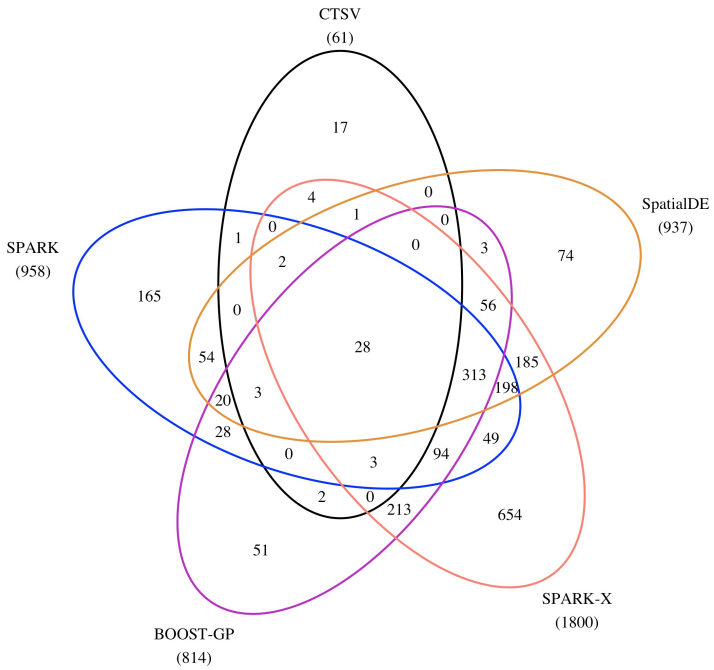

Fig. 5: Venn plot of SV genes detected by CTSV, SPARK, BOOST-GP, SPARK-X, and SpatialDE in the PDAC data. The number in the parentheses indicates the total number of SV genes detected by that method.

spots, resulting in 4,070 genes and 428 spots. We afterward applied CTSV, trendsceek (Edsgärd et al., 2018), SpatialDE (Svensson et al., 2018), SPARK (Sun et al., 2020), SPARK-X (Zhu et al., 2021), SOMDE (Hao et al., 2021), and BOOST-GP (Li et al., 2021) to the processed bulk ST data. Because trendsceek and SOMDE did not detect any SV gene in PDAC dataset, we did not display them in the downstream comparisons. The Venn plot (Figure 5) shows the SV gene overlap among CTSV, SpatialDE, SPARK, SPARK-X, and BOOST-GP. When q-value threshold is 0.05, CTSV identified 61 SV genes from 4,070 genes at the aggregated level, around a half of which were also detected by SpatialDE, SPARK, SPARK-X, and BOOST-GP. In contrast, each of the competing methods detected more than 800 SV genes. This may be because the competing methods do directly incorporate the cell type proportions and lead to false positives.

Figure 6(a)-(c) displays the overall spatial expression patterns of three representative SV genes identified by both CTSV and other methods, and we can see that although CTSV detected much smaller SV genes than other methods, it still captured important spatial patterns. The spatial expression of SV genes only detected by CTSV in Figure 6(d)-(f) indicates that CTSV is able to find some SV genes that other methods ignored despite claiming a lot of SV genes. More importantly, CTSV can recognize the SV genes in a cell-type-specific manner. For example, in

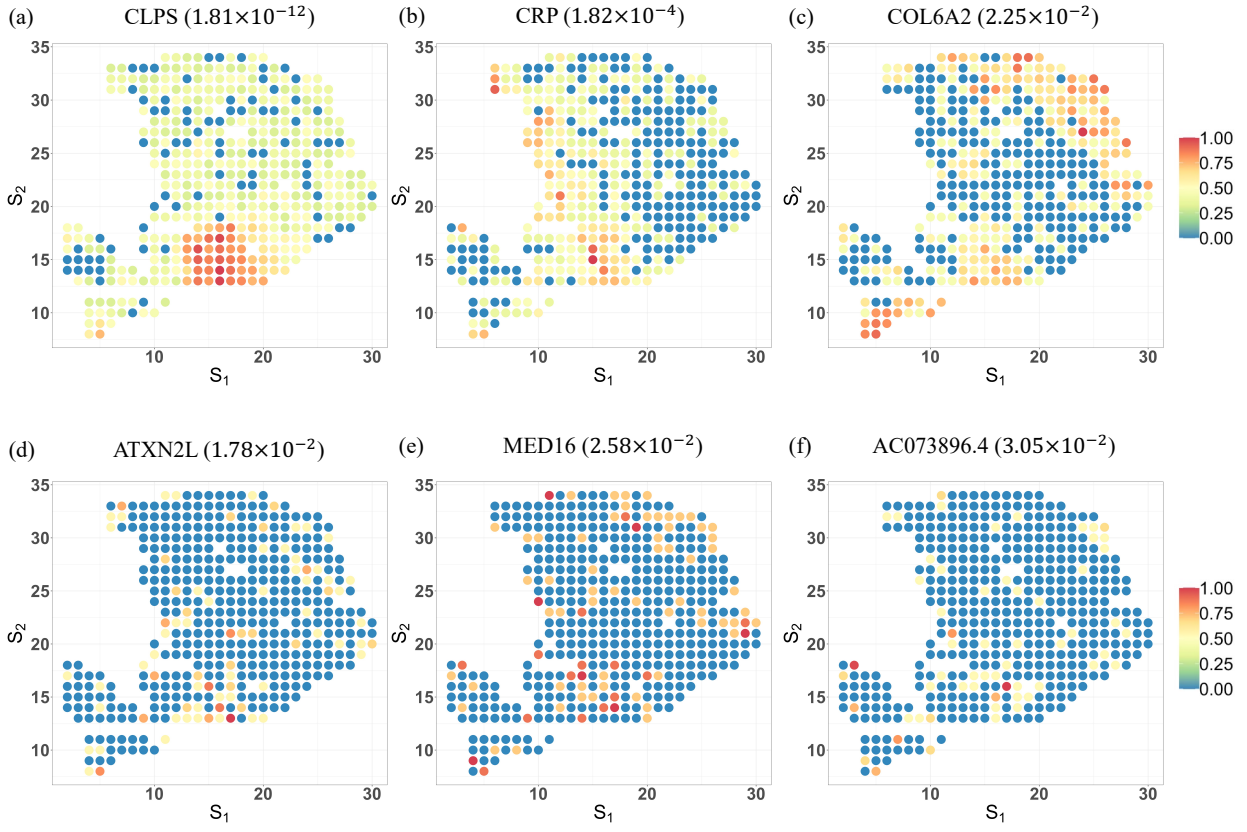

Fig. 6: Genes' spatial expression patterns in the PDAC data. (a-c) The spatial expression patterns of genes detected by all the methods—CTSV, SPARK, SPARK-X, BOOST-GP, and SpatialDE. (d-f) The spatial expression patterns of genes only identified by CTSV. The number in the parentheses is the associated q-value given by CTSV at the aggregated level.

Table 3, gene *MED16* is SV in antigen presenting ductal cells, and gene *ARHGDIB* is SV in cancer cells.

For the identification of cell-type-specific SV genes, we compared the performance between CTSV and SPARK-X. In SPARK-X, each spot was assigned to the major cell type of that spot, and then SPARK-X was applied to spots that belong to the same cell type. Table 3 shows the SV gene number in each cell type for the two methods as well as the number of overlapping SV genes. It shows that SPARK-X failed to detect any cell-type-specific SV gene except in cancer cells, while it captured too many SV genes in cancer cells.

In addition, some cell-type-specific SV genes of CTSV were found to be associated with meaningful biological functions. Table 4 displays these genes. For example, *ARHGDIB* in cancer cells, which was not identified by SPARK-X, encodes the protein RhoGDI2 that functions as a metastasis suppressor in human cancer (Gildea et al., 2002) and plays an important role in tumor dormancy regulation (Said et al., 2011). *ISG15* found in antigen presenting ductal

Table 3: Number of SV genes in eac cell type by CTSV and SPARK-X.

Cell types	CTSV	SPARK-X	overlapping genes
Antigen presenting ductal cells	13	0	0
Centroacinar ductal cells	31	0	0
High/hypoxic ductal cells	6	0	0
Terminal ductal cells	6	0	0
Cancer cells	15	673	9
Macrophages	12	0	0

cells is associated with the reinforcement of cancer stem cells' self-renewal, invasive capacity, and tumorigenic potential in PDAC (Sainz et al., 2014). In terminal ductal cells, *JADE1* may contribute to the development of pancreatic cancer (Liu et al., 2015). *CLPS* was detected as an SV gene in more than one cell type, and the pancreatic lipase requires the colipase protein encoded by *CLPS* for efficient dietary lipid hydrolysis (Lowe, 1997; Van Tilbeurgh et al., 1999). These observations illustrate that CTSV can help us gain more insights into the relationship between SV genes and diseases.

Table 4: Cell-type-specific SV genes detected by CTSV

Cell types	SV genes
Antigen presenting ductal cells	*AC092798.1*, *AL139039.2*, *CEL*, *CERS5*, *CLPS*, *CTRB1*, *CTRB2* *DUOXA2*, *FP671120.4*, *GAPDH*, *GP2*, *ISG15*, *MED16*
Centroacinar ductal cells	*AC009078.2*, *AC090114.1*, *C3*, *C4A*, *CD63*, *CD74*, *CEL*, *CELA3A* *CELA3B*, *CLPS*, *COL6A2*, *CPA1*, *CPA2*, *CPB1*, *CRP*, *CTRB1* *CTRB2*, *CTRC*, *DUOXA2*, *ELF3*, *FUT11*, *GP2*, *HEIH*, *IFI6* *IGHGP*, *KRT8*, *LCN2*, *MMP1*, *MMP14*, *MUC5B*, *NR4A1*
High/hypoxic ductal cells	*AL139039.2*, *APBB1*, *ATXN2L*, *FYCO1*, *GALNT14*, *MMP23A*
Terminal ductal cells	*AC022558.1*, *AL139039.2*, *CLPS*, *COLGALT2*, *JADE1*, *MCRIP2*
Cancer cells	*AC022558.1*, *AL139039.2*, *ARHGDIB*, *C3*, *CEL*, *CHMP6*, *CLPS*, *CLU* *CPA1*, *CPB1*, *CTRB1*, *CTRB2*, *ELN*, *GALNT14*, *LINC00685*
Macrophages	*AC073896.4*, *CBLC*, *CDKN1A*, *COLGALT2*, *DES*, *ELF3* *FGFRL1*, *FTL*, *GALNT14*, *IGFBP4*, *LNPEP*, *NSDHL*

5 Conclusion

In this paper, we developed a cell-type-specific SV gene detection method (CTSV) for bulk ST data. CTSV directly models raw count data through a zero-inflated negative binomial distribution, incorporates cell-type proportions, and relies on the R package **pscl** (Zeileis et al.,

2008) to fit the model. To capture different types of spatial patterns, five spatial effect functions are used, and then CTSV applied the Cauchy combination rule (Liu et al., 2019) to obtain p-values for robustness.

In simulation studies, CTSV was not only shown to be the most powerful approach at the aggregated level in the three spatial expression settings, but it also outperformed SPARK-X in terms of cell-type-specific SV gene detection, perhaps due to the direct consideration of cell-type proportions. In the analysis for pancreatic ductal adenocarcinoma data, CTSV also identified reasonable cell-type-specific SV genes that are related to meaningful biological functions.

Several extensions are worth exploring in the future. First, for robustness, we choose five simple spatial effect functions for h_1 and h_2, and it is better to utilize nonparametric statistical methods to directly fit the functions, such as splines or wavelets. Second, it is more helpful to incorporate prior knowledge of the tissue images (Hu et al., 2021). Third, when it comes to single-cell spatial expression data, we can also apply CTSV by setting the proportion of the cell type to which this cell belongs as one and the proportions of other cell types as zero.

6 Data availability

The PDAC datasets are publicly available in Gene Expression Omnibus with accession code GSE111672.

Acknowledgement

We thank the authors of SPOTlight (Elosua-Bayes et al., 2021) for generously providing annotated single-cell RNA-seq PDAC data. Xiangyu Luo was supported in part by National Natural Science Foundation of China (11901572) and the fund for building world-class universities (disciplines) of Renmin University of China. This research was supported by Public Computing Cloud, Renmin University of China.

References

Cable, D. M., E. Murray, L. S. Zou, A. Goeva, E. Z. Macosko, F. Chen, and R. A. Irizarry (2021). Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology*, 1–10.

Close, J. L., B. R. Long, and H. Zeng (2021). Spatially resolved transcriptomics in neuroscience. *Nature Methods 18*(1), 23–25.

Dong, R. and G.-C. Yuan (2021). SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biology 22*(1), 1–10.

Edgar, R., M. Domrachev, and A. E. Lash (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research 30*(1), 207–210.

Edsgärd, D., P. Johnsson, and R. Sandberg (2018). Identification of spatial expression trends in single-cell gene expression data. *Nature Methods 15*(5), 339–342.

Elosua-Bayes, M., P. Nieto, E. Mereu, I. Gut, and H. Heyn (2021). SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Research 49*(9), e50–e50.

Gildea, J. J., M. J. Seraj, G. Oxford, M. A. Harding, G. M. Hampton, C. A. Moskaluk, H. F. Frierson, M. R. Conaway, and D. Theodorescu (2002). RhoGDI2 is an invasion and metastasis suppressor gene in human cancer. *Cancer Research 62*(22), 6418–6423.

Hao, M., K. Hua, and X. Zhang (2021). SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. *Bioinformatics*. btab471.

Hu, J., X. Li, K. Coleman, A. Schroeder, N. Ma, D. J. Irwin, E. B. Lee, R. T. Shinohara, and M. Li (2021). 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.

Larsson, L., J. Frisén, and J. Lundeberg (2021). Spatially resolved transcriptomics adds a new dimension to genomics. *Nature Methods 18*(1), 15–18.

Li, Q., M. Zhang, Y. Xie, and G. Xiao (2021). Bayesian modeling of spatial molecular profiling data via Gaussian process. *Bioinformatics 37*(22), 4129–4136.

Liu, P., W. Jiang, Y. Han, L. He, H. Zhang, and H. Ren (2015). Integrated microrna-mrna analysis of pancreatic ductal adenocarcinoma. *Genet Mol Res 14*(3), 10288–97.

Liu, Y., S. Chen, Z. Li, A. C. Morrison, E. Boerwinkle, and X. Lin (2019). 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.

Lowe, M. E. (1997). Structure and function of pancreatic lipase and colipase. *Annual Review of Nutrition 17*(1), 141–158.

Luo, X., C. Yang, and Y. Wei (2019). Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. *Nature Communications 10*(1), 1–12.

Moncada, R., D. Barkley, F. Wagner, M. Chiodin, J. C. Devlin, M. Baron, C. H. Hajdu, D. M. Simeone, and I. Yanai (2020). Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nature Biotechnology 38*(3), 333–342.

Pillai, N. S. and X.-L. Meng (2016). An unexpected encounter with Cauchy and Lévy. *The Annals of Statistics 44*(5), 2089–2097.

Preisser, J. S., K. Das, D. L. Long, and K. Divaris (2016). Marginalized zero-inflated negative binomial regression with application to dental caries. *Statistics in Medicine 35*(10), 1722–1735.

Rahmani, E., R. Schweiger, B. Rhead, L. A. Criswell, L. F. Barcellos, E. Eskin, S. Rosset, S. Sankararaman, and E. Halperin (2019). Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature Communications 10*(1), 1–11.

Rao, N., S. Clark, and O. Habern (2020). Bridging genomics and tissue pathology: 10x genomics explores new frontiers with the visium spatial gene expression solution. *Genetic Engineering & Biotechnology News 40*(2), 50–51.

Rodriques, S. G., R. R. Stickels, A. Goeva, C. A. Martin, E. Murray, C. R. Vanderburg, J. Welch, L. M. Chen, F. Chen, and E. Z. Macosko (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science 363*(6434), 1463–1467.

Said, N., S. Smith, M. Sanchez-Carbayo, D. Theodorescu, et al. (2011). Tumor endothelin-1

enhances metastatic colonization of the lung in mouse xenograft models of bladder cancer. *The Journal of Clinical Investigation 121*(1), 132–147.

Sainz, B., B. Martín, M. Tatari, C. Heeschen, and S. Guerra (2014). ISG15 is a critical microenvironmental factor for pancreatic cancer stem cells. *Cancer Research 74*(24), 7309–7320.

Ståhl, P. L., F. Salmén, S. Vickovic, A. Lundmark, J. F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss, et al. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science 353*(6294), 78–82.

Storey, J. D., A. J. Bass, A. Dabney, and D. Robinson (2020). *qvalue: Q-value estimation for false discovery rate control*. R package version 2.22.0.

Storey, J. D. and R. Tibshirani (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences 100*(16), 9440–9445.

Sun, S., M. Hood, L. Scott, Q. Peng, S. Mukherjee, J. Tung, and X. Zhou (2017). Differential expression analysis for RNAseq using Poisson mixed models. *Nucleic Acids Research 45*(11), e106–e106.

Sun, S., J. Zhu, and X. Zhou (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods 17*(2), 193–200.

Svensson, V., S. A. Teichmann, and O. Stegle (2018). SpatialDE: identification of spatially variable genes. *Nature Methods 15*(5), 343–346.

Van Tilbeurgh, H., S. Bezzine, C. Cambillau, R. Verger, and F. Carriere (1999). Colipase: structure and interaction with pancreatic lipase. *Biochimica et Biophysica Acta (Bba)-Molecular and Cell Biology of Lipids 1441*(2-3), 173–184.

Williams, C. K. and C. E. Rasmussen (2006). *Gaussian processes for machine learning*, Volume 2. MIT press Cambridge, MA.

Zeileis, A., C. Kleiber, and S. Jackman (2008). Regression models for count data in R. *Journal of Statistical Software 27*(8), 1–25.

Zheng, S. C., C. E. Breeze, S. Beck, and A. E. Teschendorff (2018). Identification of differentially methylated cell types in epigenome-wide association studies. *Nature Methods 15*(12), 1059–1066.

Zhu, J., S. Sun, and X. Zhou (2021). 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.

Zhuang, X. (2021). Spatially resolved single-cell genomics and transcriptomics by imaging. *Nature Methods 18*(1), 18–22.