Simulator of Spatial Transcriptomics Data (SoST)

Zhentao Yu, Weifang Liu, Yun Li

September 2022

1 Abstract

As spatial transcriptomics (ST) technologies have emerged and rapidly evolved, many computational methods have been developed specifically for ST data, including but not limited to identification of spatially variable genes (SVG), spatial regions, and cell type deconvolution. Due to the lack of gold standard data, simulated datasets are commonly used to evaluate and validate these methods. However, current simulations of ST data are often constructed by synthesizing a map with "bulk" samples, pre-existing single cells, as spots. Reproducible and adjustable code or function to manifest different features of ST data is not available. Here, we present the Simulator of Spatial Transcriptomics Data (SoST) for reproducible and customizable simulation of spatially resolved gene expression data. Based on two-level gamma-Poisson distribution and adjustable geometric patterns, SoST is able to simulate spatially variable genes in terms of one gene on both spot and single cell level and mimic spatially resolved data with multiple cell types and many genes.

2 Introduction

Traditional RNA sequencing approaches such as scRNA-seq technology, provide quantitative information of single cells on gene expression but fail to take spatial information into account. However, spatially resolved transcriptomics technologies, unlike single-cell sequencing, are able to provide information of single-cell heterogeneity and identify cell types while retaining information on spatial context at the same time, which allows us to understand key properties of tumor biology, cell biology and neurobiology. For example, we can better understand molecular behaviors of single cells contained in multicellular tissue if their physical locations can be learned. It is because cells in distinct regions or microenvironments within the same tissue could differentially express a set of genes since cells have influences on and also are affected by cells in their neighborhood. The consideration of the importance of spatial context and physical position of spot/single-cell level features, usually lost in scRNA-seq and bulk data, address the technological advantage in spatial transcriptomics [1]. As ST data has become available and popular, new methods thrive and attempt to unlock its potential. The concern of these methods is different from those designed and applied for analysis of scRNA-seq data. In a single cell experiment, the gene expression profile of cells is known, and a common task is to assign cells to groups(clustering) and then apply tests for differentially expressed genes. In contrast, ST dataset either don't have information of individuals cells due to its lack of resolution (i.e., 10x Visium) or can't cover every location/spot (each usually contains 10-100 cells) in one tissue (i.e., seqFISH, seqFISH+). [1] Many existing analysis methods for ST data concentrate on two aspects. First, identification of spatially variable genes. This object is taken by tools such as SpatialDE, Trendsceek and Spark [2][3][4]. Second, deconvoluting spatial transcriptomics spots to

infer the proportion of cell types within them. This object is taken by tools such as RCTD, SPOTlight, Tangram and cell2location [5][6][7]. However, currently we do not have mature ST datasets with both high resolution and throughput to test the performance of methods on ST data. In other words, ground truth of ST data is needed to test whether an analysis method can work well or compare two methods. In this situation, another common way to test the performance of methods is through a simulated dataset. The major advantage of evaluating methods with simulated datasets is that different features, parameters and assumptions can be generated at low cost. It can be also observed that most spatial transcriptomics analysis packages are implemented and tested through simulation to address their effectiveness. However, simulation data are always specifically designed for individual methods. In this case, A reproducible and adjustable simulation framework to meet different research interests is needed. In this work we present SoST, a reusable R function for simulation of spatially transcriptomics data. SoST is a simulation framework allowing researchers to efficiently simulate ST dataset in a reproducible way. The simulated data is able to provide information of one/multiple gene expression counts of spots or single-cells, cell number and cell proportion of one/multiple cell types.

3 Methods

3.1 Simulation of a single gene

A gene is declared spatially variable if an observable change or difference in expression levels or counts between different spatial locations is statistically significant. However, in ST data, genes expressing differentially in different spots possibly reflect patterns, which can be attributed from heterogeneous distribution of cell number or proportion. These factors can confound the association between gene expression and spatial location, since cell number/proportion of one/multiple cell types can be associated with both spatial location and gene expression. For example, a gene actually is not highly expressed in cells within a spot. However, due to aggregation of a large number of cells normally expressing the gene or a large proportion of some specific cell types highly expressing the gene, the total gene expression count of that spot is high. In this case, we call the spatial pattern which is not only associated with spatial locations as a pseudo-spatial pattern. SoST implements six simulation models including Splat [8], a simulation model for scRNA-seq data. Four simulation models are aimed at simulating spatially variable genes on spot level and single-cell level in different scenarios. Two models including Splat are aimed at simulating spatially resolved data providing information of multiple cell types and multiple genes. To simulate SVG, SoST used the parameters that can be customized by the user, to generate a SVG with three predefined geometric patterns: hotspot, linear gradient and streak. The main result of this simulation step is a matrix of counts where each entry is the total gene expression counts of all cells within the spot. The notation of spatial location of the result matrix is the same as that of ST data. For example, the entry in the ith row and jth column of the m×n result matrix represents the sum of gene expression counts of all cells within the spot, whose spatial location is ith row and jth column on the real tissue (m×n grids). SoST also allows users to simulate pseudo-spatial patterns by setting parameters to define cell abundance and cell proportion. The results of pseudo-spatial pattern simulation provide information of single cells contained in spots. To simulate ST data with multiple genes and multiple cell types, SoST allows users to simulate realistic scRNA-seq data by Splat [8] and allocate those simulated single cells to locations with different density. If no ideal scRNA-seq data can be used to simulate desired cells, a synthetic ST data can still be generated by our own simulation distribution model with adjustable parameters. The main result of this simulation step is a m×n matrix where each entry represents cell-proportion of different cell types contained in the corresponded spatial spot.

3.1.1 Simulation Model for Spatially Variable Gene on Spot Level

In this simulation model, gene expression counts of each spot (sum of gene counts of all cells within each spot) is modeled by negative binomial distribution. It allows to specify desired spatial patterns at spot level. Currently it has three predefined patterns: hotspot, linear gradient and streak. One gene is considered in this simulation model [2]. Notation: Let l be the length of the pattern area. Let L be the length of the background area. Let (i, j) denote the spatial location of each spot. $i, j \in \{1, 2, \ldots, L\}$ Let $(i^*, j^*), (., j^*)$ be the spatial location of the spot/column with highest or lowest expected gene counts. We will call that spot/column as an index spot/column. Let p be the thickness of each geometric layer. Let q be the order of layers. For example, index spot is layer zero q_0; the adjacent layer of index spot is the first layer: $q = 1$; the adjacent outer layer of the first layer is second layer: $q = 2$. Let r be ratio of means of gene expression counts between two adjacent layers. r can be either less or greater than 1. Let $Z_{(i,j),q}$ be random variable representing gene expression counts of the spot which belongs to layer q and is located on ith row, j^{th} column. (i and j have different range depending on q.) Let q be mean of gene expression counts of each spot in layer q. Let be dispersion parameter of negative binomial distribution followed by $Z_{(i,j),q}$. (This one and all dispersion parameter of different simulated distributions in following sections are customizable) Let q_0 be the mean of gene expression counts of index spots. Distribution model for$Z_{(i,j),q}$:

$$Z_{(i,j),q} \sim NegativeBinomial(\mu_q, \phi) \tag{1}$$

To meet different desired geometric patterns (hotspot, linear gradient and streak) [2] numerical relationships between mean of gene expression counts of spots in layer zero and other layers are available to set. For hotspot patten:

$$\mu_q = r^q \times \mu_{q_0} \tag{2}$$

For a linear gradient pattern, the relationship between layers is defined in the same way. The difference is about q_0: layer zero q_0 represents a spot in hotspot design but a column in linear gradient design. For streak pattern, a band containing several columns has a gene expression pattern whereas spots outside the band are treated as background spots. The result of this simulation model represents gene expression spatial pattern on spot level. If the gene expression count of one spot is high, it only indicates the sum of gene counts of all cells within that spot is high. In other words, it conceals the information about single cells within spots.

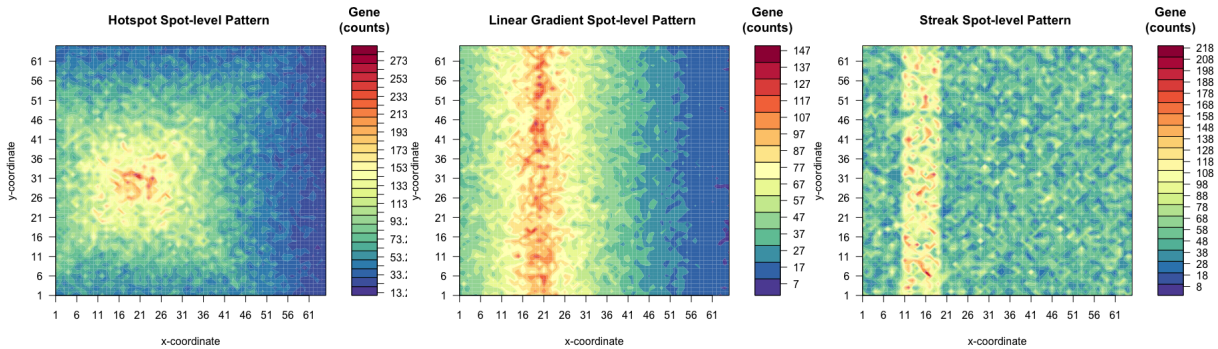

Figure 1: Examples of three types spot-level pattern:hotspot, linear gradient and streak

3.1.2 Simulation Models for Spatially Variable Gene at single-cell Level

Spatial gene pattern at spot level may be not only associated with spatial location. Information of cells such as cell abundance or proportion of different cell types within a spot can contribute to high gene expression counts. Hence, simulation models for SVG at single-cell level should take those factors into account. We consider three scenarios to simulate SVG on single-cell level: spatial pattern driven by 1.spatial location 2.cell abundance associated with spatial location 3.cell proportion associated with cell types and spatial location.

Scenario 1: Spatial Pattern Driven by Spatial Location

In this scenario, the gene spatial pattern is driven by spatial location. Let t be the index of individual cells in the spot and $t \in \{1, 2, \ldots, t_{(i,j),q}\}$. Let $T_{(i,j),q}$ be random variable representing cell number in i, j spot belonging to layer q . Let $Y_{(i,j),q}$ be random variable representing gene expression counts of cell t in i, j spot belonging to layer q. Let Λ_q be mean of gene counts of each cell in spots belonging to layer q. Let α_q be the shape parameter of gamma distribution followed by q. Let α_{q_0} be the shape parameter of gamma distribution followed by Λ_{q_0}, the expected counts of cells in index spot (i^*, j^*)belonging to layer zero q_0. Distribution of cell number $T_{(i,j),q}$:

$$T_{(i,j),q} \sim Poisson(\lambda) \tag{3}$$

To account for extra Poisson variation of gene counts data, Poisson-gamma mixture is considered:

$$Y_{(i,j),q} \sim Poisson(q), \Lambda_q \sim Gamma(q, \beta) \tag{4}$$

Gene counts of a spot is summation of counts of all cells within that spot:

$$Z_{(i,j),q} = \sum_{t=1}^{t_{(i,j),q}} Y_{(i,j),t,q} \tag{5}$$

Numerical relationships between expected mean of gene expression counts of cells within spots in layer zero and layer q is available to set:

$$\alpha_q = r^q \times \alpha_{q_0} \tag{6}$$

The main result of this simulation is a matrix where each entry represents the average gene expression counts by summing up gene counts of all cells within every spot and then dividing them into

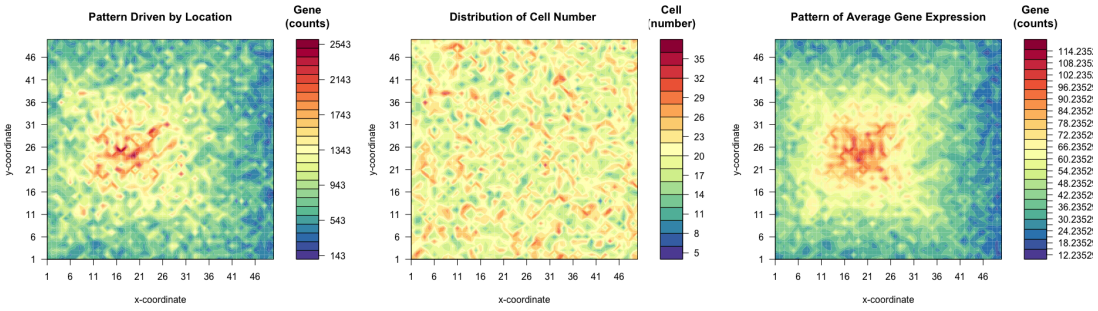

Figure 2: Example output of simulated hotspot spatial pattern driven by spatial location (Pictures are labeled as 1, 2, 3 from left to right.)

simulated cell number. This is so called the spatial pattern on single-cell level. With summation of counts of cells within every spot, we are able to get the spatial pattern on spot level.

In the example above, it is corresponded to a spatial pattern driven by spatial location (true positive) with observable spatial pattern of gene counts on spot level(graph 1), no pattern on distribution of cell number(graph 2), observable spatial pattern on average gene expression counts(graph 3).

Scenario 2: Spatial Pattern Driven by Cell Number

In this scenario, Cells in spots with higher expression are denser in space so that those spots contain more cells, but gene expression is not highly variable across cells on different spatial locations. In other words, the spatial pattern on spot level is a pseudo-spatial(false positive) pattern driven by cell number. There is an association between spatial location and cell number. Notation: Let λ_q be expected number of cells in spots belonging to layer q. Let q_0 be expected number of cells in spots belonging to layer q_0. Distribution of cell number $T_{(i,j),q}$:

$$T_{(i,j),q} \sim Poisson(q) \tag{7}$$

Relationship no longer exists between mean of gene counts but cell number in spots belonging to layer zero and layer q:

$$\lambda_q = r^q \times \lambda_{q_0} \tag{8}$$

Gene counts of cells across all locations follow the same distribution:

$$Y_{(i,j),q} \sim NegativeBinomial(\mu, \phi) \tag{9}$$

Gene counts of a spot is summation of counts of all cells within that spot:

$$Z_{(i,j),q} = \sum_{t=1}^{t_{(i,j),q}} Y_{(i,j),q} \tag{10}$$

The main result of this simulation is a matrix where every entry represents the simulated cell number of every spot. We can get the spatial pattern on spot-level by summation of gene counts of cells across all locations. This simulation actually generates a pseudo-spatial pattern since it's caused by cell number instead of highly variable gene expression of cells in those spots with high gene expression. In the example below, it is corresponded to a spatial pattern driven by cell number (false

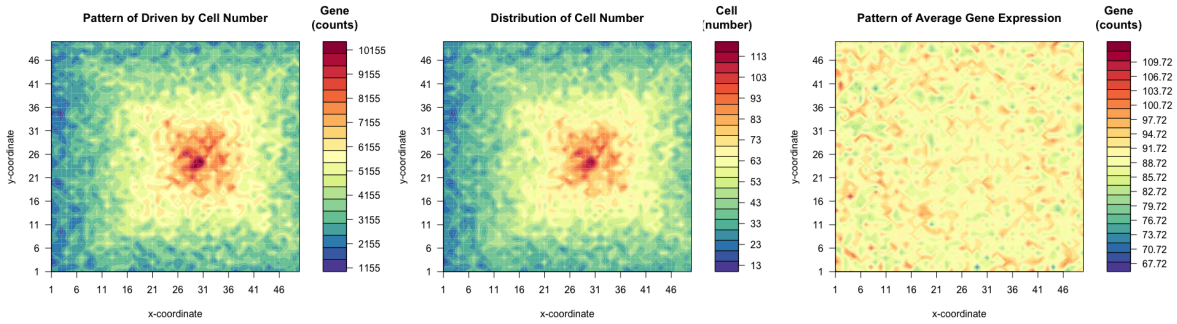

Figure 3: Example output of simulated hotspot spatial pattern driven by cell number (Pictures are labeled as 1, 2, 3 from left to right.)

positive) with observable spatial pattern on gene counts on spot level(graph 1),observable pattern of distribution of cell number(graph 2), no spatial pattern on average gene expression counts.(graph 3).

Scenario 3: Spatial Pattern Driven by Cell Proportion

In this scenario,the target gene is highly expressed in one cell type but not others. Some spots have a higher proportion of those cells that highly express the gene, and those spots observe a spatial pattern. Therefore, the spatial pattern is driven by cell type proportion(with association between cell type and cell location, cell type and average expression). $k \in \{k_1, k_2\}$. (two cell types are considered) Let $R_{(i,j),q,k}$ be proportion of cell type k in (i,j) spot belonging to layer q. Let s be ratio of cell proportion of cell type k_1 of each pair of spots between two adjacent layers. Let k be mean of gene expression counts of individual cells belonging to cell type k. Let $Y_{(i,j),q,k}$ be random variable representing gene counts of individual cell t of cell type k in (i,j) spot belonging to layer q. $t \in \{1, 2, \ldots, t_{(i,j),q,k}\}$ Let $\bar{Y}_{(i,j),q,k}$ be average gene count of cells of cell type k in (i,j) spot belonging to layer q. Assuming that only cell type k_1 exists in index spot (i_*, j_*), pattern of $R_{(i,j),q,k}$ is defined as:

$$\begin{aligned} R_{(i,j),q,k_1} &= s^q \times R_{(i^*,j^*),q_0,k_1} \\ R_{(i,j),q,k_2} &= 1 - R_{(i,j),q,k_1} \end{aligned} \tag{11}$$

Assuming that cell number across all spots have the same mean, distribution of cell number $T_{(i,j),q}$ at (i,j) spot in layer q is modeled as:

$$T_{(i,j),q} \sim Poisson(\lambda) \tag{12}$$

Distribution of gene counts $Y_{(i,j),t,k,q}$of individual cell t of cell type k in (i,j) spot and layer q is modeled as:

$$\begin{aligned} Y_{(i,j),t,k_1,q} &\sim NegativeBinomial(\mu_{k_1}, \phi) \\ Y_{(i,j),t,k_2,q} &\sim NegativeBinomial(\mu_{k_2}, \phi), \end{aligned} \tag{13}$$

where $\mu_{k_1} > \mu_{k_2}$. Then gene counts $Z_{(i,j),q}$ of (i,j) spot in layer q can calculated as:

$$Z_{(i,j),q} = \sum_{t=1}^{t_{(i,j),q}} Y_{(i,j),t,k,q} \tag{14}$$

In this way, we can define a map with observable spatial gene pattern on spot level but no pattern on average gene expression of each cell type, since the expected gene counts of cell type k_1 and

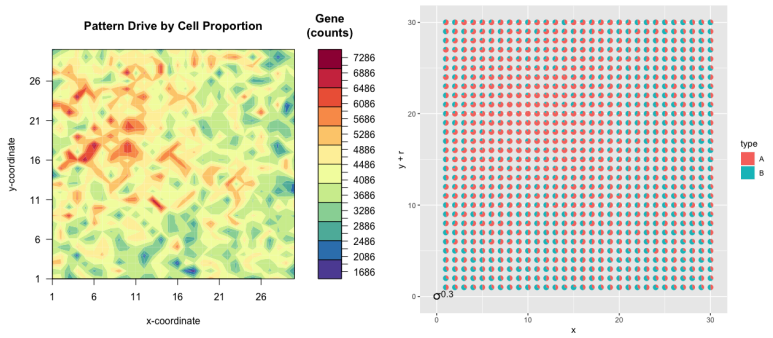

Figure 4: Example output of simulated hotspot spatial pattern driven by cell proportion

k_2 are fixed. So, the simulated spatial pattern on spot level is a pseudo-spatial pattern. Besides, the pattern of cell proportion across spots can be observed and is similar to spot-level spatial gene pattern. The main results are two matrices: 1. A matrix where every entry represents total gene counts of the corresponding spot. 2. A data frame that contains information about cell proportion of two cell types on every spot. In the example above, it is corresponded to a spatial pattern driven by cell proportion (false positive) with observable spatial pattern on gene counts on spot level(graph 1),observable spatial pattern on cell type proportion(graph 2).

3.1.3 Simulation Model for Gene Variance

Variance is another important feature of genomics data and rather vital in ST data. In this simulation model, negative binomial distribution is still considered to model gene expression counts across spots and parameters of dispersion is the interested one instead of mean. The model generates ST data with pattern on sample variance. The three predefined patterns i.e., hotspot, linear-gradient and streak, are considered to be the specified parameter of dispersion in the simulated negative binomial distribution. Distribution model for Z(i,j),q:

$$Z_{(i,j),q} \sim Negativebinomial(\mu_q, \phi_q) \tag{15}$$

The mode of pattern changing is consistent with those in the simulation model of spatially variable genes but is assigned on the parameter of dispersion ϕ. For hotspot pattern:

$$\phi_q = r^q \times q_0 \tag{16}$$

For a linear-gradient pattern, q_0 changes from index spot into column as defined in the previous. For streak pattern, the pattern of dispersion parameter is only on the index band containing the number of columns. The outside of the index band is defined with default value as background. This simulation model addresses variability pattern/distribution of gene expression. Heterogeneity of variance in gene expression profiles within a cell population can reflect functionally different cell types from sub-populations.

3.2 Simulation with multiple genes

Besides simulating individual SVG in different scenarios, SoST takes gene-gene correlation into account and provides a framework to generate ST data considering multiple cell types and genes.[9] This simulation process consists of two steps. The first step is simulating scRNA-seq data from real-data based simulation conducted by ESCO [8], a scRNA-seq data simulator using copula to impose gene-gene co-expression. The second step is to allocate generated single-cell data on spatial locations with densities. Distribution model of location and cell density is simulated based on our own models.

3.2.1 Real-data Based Simulation Model of scRNA-seq Data:

ESCO [8] simulation model captures most features observed in scRNA-seq data including zero-inflation, expression outlier gene, differential library size between cells, and mean-variance trend. It uses parametric distribution with parameters estimated from real data.

3.2.2 Simulation Model for Spatial Locations and Density

With simulated single-cell RNA data, the number of cell types within each spatial location are simulated based on different situations of sparsity and density. Notation: Let N be a random variable

representing the number of spots (locations) containing cells. Let D be mean of cell number. Let C be total number of spots in the background, which is a constant. Two types of sparsity are modeled as:

$$\begin{aligned} Uniform : N \sim Gamma(\mu = C \times 0.9, \sigma^2 = \frac{\mu}{0.3}) \\ Sparse : N \sim Gamma(\mu = C \times 0.1, \sigma^2 = \frac{\mu}{0.3}) \end{aligned} \tag{17}$$

Cell number T in those spots (locations) containing numbers is modeled as:

$$T \sim Poisson(D) \tag{18}$$

Two types of cell density are modeled as:

$$\begin{aligned} HighDensity : D \sim Gamma(\mu_1, \sigma^2 = \mu_1) \\ LowDensity : D \sim Gamma(\mu_2, \sigma^2 = \mu_2) \end{aligned} \tag{19}$$

,

where $\mu_1 > \mu_2$. For every cell type, N locations are chosen from all C locations by random. For each chosen location, cell number T is simulated by sampling from Poisson distribution with rate parameter D. The cell number of other locations is 0. In default setting, $\mu_1 = 2.5$ and $\mu_2 = 0.8$ [10]. At last, we can randomly allocate cells of each cell type to those simulated locations with designed density. The main result of this simulation is a data frame providing information about cell proportion and abundance on each spot (locations).

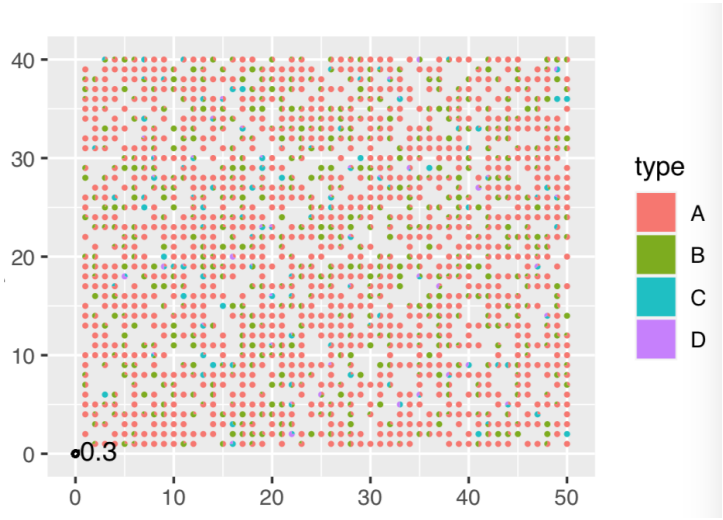

Figure 5: Simulated ST data with information of cell proportion. Cell type A: Uniform and high density. Cell type B: Uniform and low density. Cell type C:Sparse and high density. Cell type D: Sparse and low density