Bioprocess 生物过程, 2023, 13(1), 57-63
Published Online March 2023 in Hans. https://www.hanspub.org/journal/bp
https://doi.org/10.12677/bp.2023.131008

Hans汉斯

用于空间分辨转录组学数据分析的统计方法

王　琳，赵桂华

上海理工大学理学院，上海

收稿日期：2023年2月23日；录用日期：2023年3月17日；发布日期：2023年3月29日

摘　要

近年来，空间转录组学的发展使得对细胞转录组及其空间位置进行多重分析得以实现。伴随着实验技术能力与效率的日益提升，发展分析方法的要求也逐渐显现。生成空间分辨转录组(SRT, Spatially Resolved Transcriptome)数据的技术正在迅速改进，并应用于研究各种生物组织。研究空间定位基因表达如何为不同组织发育提供新的见解是至关重要的。我们回顾了用于分析不同SRT数据集的可用包，重点是识别空间可变基因(SVGs, Spatially Variable Genes)。另外，在测序方案不断开发的过程中，有必要对现有分析方法中的基本假设进行重新评价与调整，以便使用越来越复杂的数据。为启发和协助今后模型开发工作，这里将对空间转录组学统计学习方法研究新进展进行综述，归纳出有用资源并介绍今后的挑战与机遇。

关键词

基因表达模式，空间分辨转录组，统计分析

Statistical Methods for Spatially Resolved Transcriptomic Data Analysis

Lin Wang, Guihua Zhao

College of Science, University of Shanghai for Science and Technology, Shanghai

Received: Feb. 23rd, 2023; accepted: Mar. 17th, 2023; published: Mar. 29th, 2023

Abstract

In recent years, the development of spatial transcriptomics has enabled multiple analyses of cell transcriptome and its spatial location. With the increasing ability and efficiency of experimental technology, the requirement of developing analytical methods has gradually emerged. Techniques for generating Spatially Resolved Transcriptome (SRT) data are rapidly improving and being applied to study a variety of biological tissues. It is critical to study how spatially localized gene ex-

文章引用: 王琳, 赵桂华. 用于空间分辨转录组学数据分析的统计方法[J]. 生物过程, 2023, 13(1): 57-63.
DOI: 10.12677/bp.2023.131008

pression provides new insights into different tissue development. This paper reviews the packages available for analysis of different SRT data sets, with emphasis on the identification of Spatially Variable Genes (SVGs, Spatially Variable Genes). In addition, as sequencing protocols continue to be developed, it is necessary to reevaluate and adjust the basic assumptions in existing analytical methods in order to use increasingly complex data. In order to inspire and assist future model development, this paper reviews the recent progress of statistical learning methods in spatial transcriptomics, summarizes useful resources, and introduces future challenges and opportunities.

Keywords

Gene Expression Patterns, Spatially Resolved Transcriptome, Statistical Analysis

1. 引言

近十年来，空间转录组学技术的迅速发展促进了生物学在不同领域的发现[1]。空间分辨转录组学允许我们以细胞组织为背景，对细胞转录组进行研究。空间信息这一附加维度已经表现出有效性，这为我们在细胞转录组方面的研究提供了新的视角。与此同时，空间转录组学技术的发展也加大了数据量与复杂性，给数据分析提出了新挑战。近期计算方法的发展，为分析高维数据创造了新的有效范式，例如单细胞 RNA 高通量测序(scRNA-seq, signel cell RNA-sequencing)的研究[2]。类似地，空间转录组学数据分析方法发展领域已经取得长足的进步[3] [4]。在理论方面，很多针对 scRNA-seq 数据分析而发展起来的计算方法都能够应用到空间转录组学数据的研究中[5] [6]。但是，为了充分利用空间信息，设计新的方法仍然是有必要的。

scRNA-seq 和空间转录组学数据是相辅相成的，各有其独特的性质和优势。尽管在细胞制备过程中丢失了有关细胞空间位置的信息，丢失的信息有可能通过利用细胞的基因表达模式进行重建。另一方面，空间转录组学保留空间信息，但大多数数据既不是转录组范围的广度，也不是深度的细胞分辨率。例如，当测序捕获位置大于单个细胞时，在该捕获位置测量的基因表达将来自多个细胞的混合物。为了解决这个问题，我们可以采用迁移学习的想法，它利用从类似领域学习的知识，在这些领域，数据更容易获取或更好地标记出来[7]。在实际应用中，利用从 scRNA-seq 中提取的表达谱以及从空间转录组中提取的空间模式可以实现两类数据间的知识传递，从而有助于对这两类数据进行解析。

本综述的主要目的是介绍目前用于空间分辨转录组学的统计分析方法。空间转录组学数据分析工作流程通常包括多个阶段。第一步是数据预处理，通常包括质量控制、基因表达归一化、降维、细胞类型标注等。人们可以通过空间分解、基因插补和标签转移进一步提高数据的丰富程度[8]。接下来，人们可以通过空间聚类和局部基因表达模式发现从数据中获得生物学见解，这可以进一步促进空间可变基因的识别。此外，空间转录组学数据可用于帮助重建 scRNA-seq 数据中的空间位置。尽管目前一些计算方法在该工作流程中得到了成功应用，但仍迫切需要开发更复杂的模型来应对空间转录组学数据分析中日益严峻的挑战。为了弥补不断发展的实验技术与当前计算技术之间的差距，我们在此调查了统计建模在空间转录组学中的应用，并根据应用领域将其分为识别空间表达模式基因和聚类分析。最后，我们概述了

空间转录组学方法开发领域的挑战和未来机遇。

2. 识别具有空间表达模式的统计分析方法

基于已知的细胞空间坐标及其基因表达水平的统计建模的方法为阐明空间基因表达异质性提供了统计框架。首先，输入细胞的基因表达谱和位置信息。根据输入信息，构建统计框架，明确基因表达值与细胞空间位置的依赖关系。随后，通过不同的统计方法确定显着的 SVG。接下来将介绍这些统计分析方法(表 1)。

Table 1. Computational methods for recognizing SVGs based on statistical modeling
表 1. 基于统计建模的识别 SVGs 的计算方法

Method	Platform	URL	Reference
trendsceek	R	https://github.com/edsgard/trendsceek	Edsgard *et al.*
SpatialDE	Python	https://github.com/Teichlab/SpatialDE	Svensson *et al.*
SPARK	R	https://xzhoulab.github.io/SPARK/	Sun *et al.*
SPARK-X	R	https://xzhoulab.github.io/SPARK/	Zhu *et al.*
GPcounts	Python	https://github.com/ManchesterBioinference/Gpcounts	BinTayyash *et al.*
BayesSpace	R	https://github.com/edward130603/BayesSpace	Zhao *et al.*

2.1. Trendsceek

Trendsceek [9]使用标记点过程来模拟基因表达和细胞坐标之间的关联。Trendsceek 将每个点表示为一个细胞，并将该点的标记表示为基因表达值，并计算点之间的距离。评估点的空间分布与其相关标记(表达式级别)之间是否存在显著依赖关系。对于特定距离，评估基因的标记是否取决于点的位置，换句话说，是否发生了标记分离。使用四种类型的依赖性评估方法来测试标记分离(均值标记函数，方差标记函数，标记变异函数和 Stoyan 标记相关函数)。如果分数和分数分布是相关的，则分数在不同的距离上应该是可变的。作为标记分离汇总统计信息，Trendsceek 使用了四个以前已知的两点空间统计信息。这四种统计数据都计算所有点对的汇总统计数据，以它们之间的距离 r 为条件。下文描述了四种使用的统计数据，对于每个点，P 表示距离为 r 的点对的集合，m_1，m_2 表示两个点的标记。

1) 均值标记函数，它是所有点的算术平均值，属于由距离 r 隔开的点对：

$$E_{mark}(r)=\frac{E(m_1m_2)_P(r)}{2}$$

2) 方差标记函数，这是对分离条件下的方差：

$$V_{mark}(r)=E\left[\left(m_1-E(m_1)_P(r)\right)^2\right]_P(r)$$

3) 标记变异函数标记点过程的标记变异函数，它是基于距离 r 的点对之间的标记差的平方：

$$\gamma(r)=E\left[\frac{1}{2}(m_1-m_2)^2\right]_P(r)$$

4) Stoyan 标记相关函数，它使用距离 r 处所有对标记的平方几何均值，用所有点的平方均值归一化，无论距离如何：

$$\rho(r)=\frac{E(m_1m_2)_P(r)}{\bar{m}^2}.$$

2.2. SpatialDE

SpatialDE [10]是一种基于高斯过程回归的方法。高斯过程(GP, Gaussian Process)是一个随机过程，也称为正态分布。它允许非线性回归和量化测量过程与潜在功能之间的关联，并且 GP 对于 SRT 数据可用于模拟基因表达的空间梯度变化。SpatialDE 建立了具有高斯核的基因表达谱的线性混合模型，并将每个基因的变异分解为空间或非空间变异。使用观测噪声对非空间变异进行建模，空间变异由基因表达值和空间细胞坐标的协方差矩阵表示。对于每个高斯核，SpatialDE 使用似然检验与零模型进行比较计算近似 p 值，并识别具有显着空间变异性的基因。

SpatialDE 使用以下形式的多元正态模型对给定基因在空间坐标 $X=(x_1,x_2,\cdots,x_N)$ 上的基因表达谱 $y=(y_1,y_2,\cdots,y_N)$ 进行建模，如公式(1)所示

$$P(y|\mu,\sigma_s^2,\delta,\Sigma)=N(y|\mu\cdot 1,\sigma_s^2\cdot(\Sigma+\delta\cdot I)) \tag{1}$$

固定影响 $\mu\cdot 1$ 表示平均表达水平，而 Σ 表示基于单元格对之间的输入坐标定义的空间协方差矩阵。该模型使用的平方指数协方差函数 Σ 定义如公式(2)

$$\Sigma_{i,j}=k(x_i,x_j)=\exp\left(-\frac{|x_i-x_j|^2}{2\cdot l^2}\right) \tag{2}$$

然后通过最大边际似然来拟合模型参数，如公式(3)所示

$$LL=-\frac{1}{2}\cdot N\cdot\log(2\pi)-\frac{1}{2}\log(|\sigma_s^2\cdot(\Sigma+\delta\cdot I)|)-\frac{1}{2}\cdot(y-\mu\cdot 1)^{\mathrm{T}}(\sigma_s^2\cdot(\Sigma+\delta\cdot I))^{-1}(y-\mu\cdot 1) \tag{3}$$

为了估算统计显著性，将拟合该模型的模型可能性与对应与无空间协方差的零假设的模型的可能性进行比较，如公式(4)所示

$$P(y|\mu,\sigma^2)=N(\mu\cdot 1,\sigma^2\cdot I) \tag{4}$$

然后使用贝叶斯信息准则进行比较，如公式(5)所示

$$BIC=\log(N)\cdot M-2\cdot LL \tag{5}$$

这里的 M 表示给定模型的超参数数量，N 表示样本数量，LL (公式(3))是数据的对数边际似然。

$Y=(y_1,y_2,\cdots,y_G)$ 是表达矩阵 G 在每一个空间位置，$\mu=\{\mu_1,\cdots,\mu_k\}$ 是矩阵 K 底层的模式。所以向量 μ_k 表示模式 k。此外，令 Z 为二进制的指示矩阵，如果 $z_{g,k}=1$，则将基因 g 分配给模式 k。那么，所有基因的完整模型如公式(6)所示：

$$\begin{aligned}
&P(Y,\mu,Z,\sigma_e^2,\Sigma)=P(Y|\mu,Z,\sigma_e^2)\cdot P(\mu|\Sigma)\cdot P(Z)\\
&P(Y,\mu,Z,\sigma_e^2)=\prod_{K=1}^{K}\prod_{g=1}^{G}N(y_g|\mu_k,\sigma_e^2)\\
&P(\mu|\Sigma)=\prod_{k=1}^{K}N(\mu_k|0,\Sigma)\\
&P(Z)=\prod_{k=1}^{K}\prod_{g=1}^{G}\left(\frac{1}{K}\right)^{(z_g,k)}
\end{aligned} \tag{6}$$

2.3. SPARK

SPARK [11]通过广义空间线性模型直接对从各种空间分辨转录组学技术生成的技术数据进行建模。它依赖于用于可扩展计算的惩罚拟似然算法[12] [13]和最近开发的用于假设检验的统计公式[14]，提供对I类错误的有效控制并产生较高的统计能力。

与SpatialDE相比，SPARK进行了一些特定的改进。SPARK基于具有多个空间核(包括高斯核和周期核)的空间广义线性混合模型识别SVG，以直接对空间计数数据进行建模。为了适应不同的空间模式，SPARK默认使用十个空间内核，包括最常见的空间表达模式。SPARK通过惩罚拟似然算法估计参数，SPARK依赖于混合χ^2分布准确检验每个空间核的p值，然后利用柯西组合法则对所有p值进行组合，以得到校准良好的p值，可以有效控制I型误差的发生。此外，还开发了SPARK的高斯版本，在面对高计数的SRT数据时，可以保持稳定的性能。在SPARK方法中，对n个空间样本上的焦点基因的表达计数数据采用GLSM建模：

$$y_i\left(s_i\right) \sim \mathrm{Poi}\left(N_i\left(s_i\right)\lambda_i\left(s_i\right)\right),\ i=1,2,\cdots,n$$

$$\log\left(\lambda_i\left(s_i\right)\right)=x_i\left(s_i\right)^{\mathrm{T}}\beta+b_i\left(s_i\right)+\varepsilon_i$$

$$b(s)=\left(b_1\left(s_1\right),b_2\left(s_2\right),\cdots,b_n\left(s_n\right)\right)^{\mathrm{T}} \sim \mathrm{MVN}\left(0,\tau_1 K(s)\right)$$

$$H_0:\tau_1=0$$

其中$y_i\left(s_i\right)$为第i个样本处焦点基因的基因表达计数，$N_i\left(s_i\right)$表示第i个样本中所有基因的总计数，$\lambda_i\left(s_i\right)$是一个未知的泊松速率参数，表示第i个样本中焦点基因的潜在基因表达水平，s_i表示第i个样本的空间坐标。为了探索基因的空间表达模式，将$\log\left(\lambda_i\left(s_i\right)\right)$建模为$x_i\left(s_i\right)^{\mathrm{T}}\beta$，$b_i\left(s_i\right)$，$\varepsilon_i$三项的线性组合，$x_i\left(s_i\right)$表示样本解释变量的向量，包含一个表示截距的标量和观察到的样本解释变量。这些解释变量在分析过程中进行了调整，可能包含批处理信息、细胞周期信息或其他重要信息。β是k维回归系数，$b_i\left(s_i\right)$是一个零均值的平稳高斯过程，$\varepsilon_i\left(1\le i\le n\right)$是均值为0、方差为$\tau_2$的独立同分布的正态随机变量。由于$b_i\left(s_i\right)$的分布可知，$b(s)$服从均值为0、协方差矩阵为$\tau_1$K(s)的多元正态分布。其中$\tau_1$是缩放因子，$K_{ij}(s)=K\left(s_i,s_j\right)$是空间位置$s=\left(s_1,\cdots,s_n\right)^{\mathrm{T}}$的核函数。

2.4. SPARK-X

SPARK-X [15]是处理大型和稀疏SRT数据时SPARK的有效补充。SPARK-X建立在强大的协方差测试框架之上，基于非参数建模，SPARK-X有效降低了内存需求和计算时间，同时保持了模型可靠性。

SPARK-X基于空间位置为所有样本构建一个距离协方差矩阵$E=y\left(y^{\mathrm{T}}y\right)^{-1}y^{\mathrm{T}}$，还基于空间位置为所有样本构建了一个距离协方差矩阵，作为$\sum=S\left(S^{\mathrm{T}}S\right)^{-1}S^{\mathrm{T}}$。其中$y=\left(y_1\left(s_1\right),\cdots,y_n\left(s_n\right)\right)^{\mathrm{T}}$，$S=\left(s_1^{\mathrm{T}},\cdots,s_n^{\mathrm{T}}\right)^{\mathrm{T}}$，$y_i\left(s_i\right)$为第i个样本的基因表达计数，$s_i$表示第i个样本的空间坐标，$i\in(1,\cdots,n)$。对于这两个矩阵SPARK-X将他们居中为$\boldsymbol{E}_C=\boldsymbol{HEH}$和$\sum_C=\boldsymbol{H}\sum\boldsymbol{H}$，其中$\boldsymbol{H}=\left(\mathbf{I}-\mathbf{1}_n\mathbf{1}_n^{\mathrm{T}}/n\right)$，$I$是$n*n$的单位矩阵，$\mathbf{1}_n$是元素为1的$n$向量。SPARK-X构造以下检验统计量：$T=\mathrm{trace}\left(\mathbf{E}_C\sum_C\right)/n$。直观地说，任一协方差矩阵中的每个元素根据位置对于平均值的协调偏差来测量位置对之间的相似性。

2.5. GPcounts

GPcounts [16]利用高斯过程回归方法，该方法实现负二项式似然模型(有时为零膨胀负二项式[ZINB, Zero Inflated Negative Binomial])来对SRT数据进行建模，在处理计数数据时实现比高斯似然函数更好的拟合。负二项式似然变异的平均值基于对数链接函数建模。GPcounts提供单样本和双样本测试，以推断

空间计数数据中跨空间的差异表达基因。在单样本检验中，原假设是高斯模型，基因表达没有空间变异性，细胞之间也没有协方差。在双样本测试中有两个零假设：1) 两种条件下的基因表达没有差异；2) 构建三个 GP 后，每个样本使用一个 GP，剩余的 GP 在两个样本之间共享。GPcounts 实现 χ^2 分布以评估每个基因的 p 值。此外，GPcounts 可以使用 ZINB 模型而不是负二项式模型来处理包含太多零的数据。GPcounts，可用于建立空间或时间的大规模 scRNA-Seq 数据模型，通过使用负二项式(NB, Negative Binomial)似然对计数数据进行建模。与高斯似然模型相比，NB 似然模型应更准确地捕捉基因表达数据的分布，因为它考虑到了可能的异方差噪声和许多零计数的存在，但需要应用唯一分子标识符(UMI, Unique Molecular Identifiers)规范化[17] [18]。GPcounts 的主要目的不是识别 SVG，它还能够识别差异表达基因，执行伪时间推断，然后识别分支基因并发现时间轨迹，与大多数软件包相比，它的范围更广。

2.6. BayesSpace

最近引入了一种称为贝叶斯空间的完全贝叶斯统计方法，以基于来自空间邻域的信息提高 SRT 数据的分辨率，并进行空间聚类分析以推断具有相似基因表达模式的簇。贝叶斯空间克服了有效利用空间信息进行基因表达数据聚类的局限性和原始数据的有限分辨率。BayesSpace [19]是一种完全贝叶斯统计方法，它使用来自空间邻域的信息来增强空间转录组数据的分辨率并进行聚类分析。

3. 小结与展望

近年来，SRT 技术不断发展，成为疾病研究的新范式。越来越多的证据表明，组织中细胞的空间位置和基因表达水平之间的关联在疾病中起着关键作用，特别是在肿瘤机制和微环境中[20] [21]。鉴定具有空间可变表达模式的基因是 SRT 数据分析的关键任务，反映了相邻细胞、位置特异性状态或迁移到特定组织的细胞之间的通信。它在识别与特定组织区域相关的肿瘤标志物以及靶向治疗方面提供了广泛的适用性，探索了与特定功能相关的空间表达模式，并为肿瘤异质性的起源提供了见解。

参考文献

[1] Ståhl, P.L., Salmén, F., Vickovic, S., *et al*. (2016) Visualization and Analysis of Gene Expression in Tissue Sections by Spatial Transcriptomics. *Science*, **353**, 78-82. https://doi.org/10.1126/science.aaf2403

[2] Kharchenko, P.V. (2021) The Triumphs and Limitations of Computational Methods for scRNA-seq. *Nature Methods*, **18**, 723-732. https://doi.org/10.1038/s41592-021-01171-x

[3] Lein, E., Borm, L.E. and Linnarsson, S. (2017) The Promise of Spatial Transcriptomics for Neuroscience in the Era of Molecular Cell Typing. *Science*, **358**, 64-69. https://doi.org/10.1126/science.aan6827

[4] Dries, R., Chen, J., Del Rossi, N., *et al*. (2021) Advances in Spatial Transcriptomic Data Analysis. *Genome Research*, **31**, 1706-1718. https://doi.org/10.1101/gr.275224.121

[5] Kharchenko, P.V., Silberstein, L. and Scadden, D.T. (2014) Bayesian Approach to Single-Cell Differential Expression Analysis. *Nature Methods*, **11**, 740-742. https://doi.org/10.1038/nmeth.2967

[6] Vu, T.N., Wills, Q.F., Kalari, K.R., *et al*. (2016) Beta-Poisson Model for Single-Cell RNA-seq Data Analyses. *Bioinformatics*, **32**, 2128-2135. https://doi.org/10.1093/bioinformatics/btw202

[7] Weiss, K., Khoshgoftaar, T.M. and Wang, D. (2016) A Survey of Transfer Learning. *Journal of Big Data*, **3**, 1-40. https://doi.org/10.1186/s40537-016-0043-6

[8] Zeng, Z., Li, Y.W., Li, Y.M., *et al*. (2022) Statistical and Machine Learning Methods for Spatially Resolved Transcriptomics Data Analysis. *Genome Biology*, **23**, 1-23. https://doi.org/10.1186/s13059-022-02653-7

[9] Edsgärd, D., Johnsson, P. and Sandberg, R. (2018) Identification of Spatial Expression Trends in Single-Cell Gene Expression Data. *Nature Methods*, **15**, 339-342. https://doi.org/10.1038/nmeth.4634

[10] Svensson, V., Teichmann, S.A. and Stegle, O. (2018) SpatialDE: Identification of Spatially Variable Genes. *Nature Methods*, **15**, 343-346. https://doi.org/10.1038/nmeth.4636

[11] Sun, S., Zhu, J. and Zhou, X. (2020) Statistical Analysis of Spatial Expression Patterns for Spatially Resolved Tran-

scriptomic Studies. *Nature Methods*, **17**, 193-200. https://doi.org/10.1038/s41592-019-0701-7

[12] Breslow, N.E. and Lin, X.H. (1995) Bias Correction in Generalized Linear Mixed Models with a Single-Component of Dispersion. *Biometrika*, **82**, 81-91. https://doi.org/10.1093/biomet/82.1.81

[13] Sun, S.Q., *et al*. (2019) Heritability Estimation and Differential Analysis of Count Data with Generalized Linear Mixed Models in Genomic Sequencing Studies. *Bioinformatics*, **35**, 487-496. https://doi.org/10.1093/bioinformatics/bty644

[14] Liu, Y.W., *et al*. (2019) ACAT: A Fast and Powerful P Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *The American Journal of Human Genetics*, **104**, 410-421. https://doi.org/10.1016/j.ajhg.2019.01.002

[15] Zhu, J., Sun, S. and Zhou, X. (2021) SPARK-X: Non-Parametric Modeling Enables Scalable and Robust Detection of Spatial Expression Patterns for Large Spatial Transcriptomic Studies. *Genome Biology*, **22**, 1-25. https://doi.org/10.1186/s13059-021-02404-0

[16] BinTayyash, N., Georgaka, S., John, S.T., *et al*. (2021) Non-Parametric Modelling of Temporal and Spatial Counts Data from RNA-seq Experiments. *Bioinformatics*, **21**, 3788-3795. https://doi.org/10.1093/bioinformatics/btab486

[17] Svensson,V. (2020) Droplet scRNA-seq Is Not Zero-Inflated. *Nature Biotechnology*, **38**, 147-144. https://doi.org/10.1038/s41587-019-0379-5

[18] Townes, F.W., *et al*. (2019) Feature Selection and Dimension Reduction for Single-Cell RNA-Seq Based on a Multinomial Model. *Genome Biology*, **20**, 1-16. https://doi.org/10.1186/s13059-019-1861-6

[19] Zhao, E., Stone, M.R., Ren, X., *et al*. (2021) Spatial Transcriptomics at Subspot Resolution with BayesSpace. *Nature Biotechnology*, **11**, 1375-1384. https://doi.org/10.1038/s41587-021-00935-2

[20] Saviano, A., Henderson, N.C. and Baumert, T.F. (2020) Single-Cell Genomics and Spatial Transcriptomics: Discovery of Novel Cell States and Cellular Interactions in Liver Physiology and Disease Biology. *Journal of Hepatology*, **73**, 1219-1230. https://doi.org/10.1016/j.jhep.2020.06.004

[21] Smith, E.A. and Hodges, H.C. (2019) The Spatial and Genomic Hierarchy of Tumor Ecosystems Revealed by Single-Cell Technologies. *Trends in Cancer*, **5**, 411-425. https://doi.org/10.1016/j.trecan.2019.05.009