arXiv:2603.09061v1 [stat.AP] 10 Mar 2026

空间转录组学中空间变异基因的无分布筛选

ChanghuWang *Qiyun Huang*Zihao Chen *
北京大学数学科学学院 JinLiu †香港中文大学（深圳）数据科学学院 Ruibin Xi †北京大学数学科学学院、统计科学中心

摘要

空间转录组学（ST）技术能够在保留空间分辨率的同时进行全转录组范围的基因表达谱分析，为揭示复杂空间结构带来了前所未有的机会。由于空间转录组数据具有超高维特性，识别与未知空间簇相关的空间变异基因（SVGs）已成为空间转录组数据分析的一项核心任务。在此，我们开发了一种无分布的空间变异基因筛选方法，该方法基于一种新颖的拟似然比统计量——MM检验，并结合 knockoff 方法控制错误发现率（FDR）。MM检验利用有关未知空间域的辅助信息（例如空间距离）进行空间变异基因筛选。值得注意的是，除了适用于二维空间转录组数据集外，MM检验也非常适用于日益常见的三维（三维）、多切片空间转录组数据集。基于模拟和 34 个真实空间转录组数据集开展的大量基准测试表明，MM检验始终优于现有的空间变异基因检测方法。在一个三维小鼠脑数据集中，MM检验准确地刻画了其他方法难以识别的精细尺度结构，例如海马角锥体层和齿状回的三维结构。本文还建立了理论保证，包括选择一致性、错误发现率控制，以及后选择聚类的误差界。

关键词：多切片空间转录组学、空间变异基因、异质性检验、优化最小化算法、基于knockoff的错误发现率控制

*贡献相同
†共同通讯作者

1 引言

空间转录组学（ST）技术的出现彻底革新了基因表达分析，使研究人员能够在保留关键信息——空间信息的同时，全面开展转录组分析（Marx，2021 年）。这些尖端技术能够同时测量数千个组织位置（“空间点”）的全转录组表达，从而揭示此前难以探知的基因表达模式与组织结构之间的关系。通过保留空间上下文，空间转录组学为细胞组织结构与功能提供了前所未有的见解，其应用正在改变发育生物学、神经科学和肿瘤学研究（Tang 等，2026）。空间转录组学技术的主要分析目标是识别空间域——即具有不同细胞组成、结构模式和功能作用的组织区域。此类空间域主要通过对高维空间转录组数据进行聚类分析来界定（Dries 等，2021 年；Hu 等，2021 年）。然而，由于大量基因缺乏空间相关性，出现了显著的方法学挑战。这些与空间无关的基因会引入统计噪声，可能掩盖潜在的生物学信号，并降低聚类效率。因此，开发稳健的特征选择方法，系统识别构成空间域结构的关键基因，对于准确表征组织至关重要。

理想情况下，空间转录组数据的特征选择方法（也称为空间变异基因（SVGs）检测）应能够应对以下统计和计算方面的挑战：

- 由于缺乏有关组织空间域的先验知识，因此有必要开发无监督方法，仅依据基因固有的空间表达模式来识别空间变异基因。

- 空间转录组数据具有超高维特性，形成了大海捞针般的情形：只有一小部分基因呈现具有生物学意义的空间变异。

- 空间转录组学数据具有计数型、零膨胀和过度离散的特性，因此需要采用专门的统计建模方法，以区分真实的生物学变异与技术伪影（Sun 等，2020 年）。

- 随着能够捕获完整组织或器官结构的三维（3D）空间转录组数据集日益增多（Schott 等，2024年；Chen 等，2022年），亟需能够在三维空间中联合识别一致的空间变异基因的整合分析框架。

- 有效控制错误发现率对于在高维环境中进行可靠推断，以及对空间变异基因开展下游功能分析（如基因本体或通路分析）至关重要。

应对这些挑战需要开发稳健的统计方法，将空间模式检测与严格的特征选择原则相结合。

近年来，空间变异基因筛选在生物信息学和计算生物学文献中受到了广泛关注（Chen 等人，2025 年）。这些方法要么利用多种空间相关性指标识别空间变异基因（Sun 等人，2020 年；Zhu 等人，2021 年；Yuan 等人，2024年），要么采用空间广义线性混合效应模型，检测具有非零随机空间效应的空间变异基因（Svensson 等人，2018 年；Sun 等人，2020 年；Li 等人，2021 年）。然而，这些方法大多针对单切片分析设计，无法整合多个三维组织切片中的信息，而这在现代空间转录组数据集中日益重要。此外，这些方法通常缺乏理论保证，

例如选择一致性或错误发现率控制。此外，这些方法并非旋转不变，这意味着空间坐标的不同旋转可能导致结果不一致（SuandCui，2025 年）。

在统计学文献中，空间变异基因筛选可视为聚类分析中的特征筛选问题。特征筛选的理论基础最初由Fan and Lv (2008年在线性回归背景下确立，推动了监督学习领域方法的大量发展（Liu 等，2015年）。然而，聚类问题的无监督对应方法仍鲜有研究，尽管其在高维应用中的重要性日益凸显，尤其是在空间转录组数据中检测空间变异基因方面。在现有方法中，一类重要方法依据特征的分布模态进行评估（Chanand Hall，2010年；Jin and Wang，2016年；Banerjee 等，2017年；Wang 等，2025 年）：单峰分布表明特征与簇无关，多峰分布则说明特征与簇相关。最近，Liu 等(2022年) 提出了特征选择谱聚类（SCFC），该方法通过衡量特征与初步聚类分配之间的相关性来量化特征相关性。这些筛选方法通常具有完善的理论保证。然而，它们未纳入空间或基于图像的上下文，而这对于检测空间转录组数据中的空间域至关重要；此外，许多方法也未提供任何错误发现率控制。

为应对空间转录组数据中特征选择的挑战并促进区域检测，我们提出了一种 MM 检验统计量，通过准似然框架利用均值-方差关系，评估不同簇之间的均值齐性模型（见第 3 节）。该检验统计量通过最大化异质与齐性均值模型之间的准似然差异，并采用主化 - 最小化（MM）算法推导得出（Hunter and Lange，2004年）。MM 检验还纳入了辅助信息—

—通过距离矩阵利用空间或结构特征捕捉样本关系。该矩阵为局部均值估计和离散度建模提供依据，使 MM 检验能够适应潜在的空间结构，并提高识别聚类相关特征的灵敏度。随后，通过保留 MM 检验统计量超过预设阈值 t_n 的变量来进行特征选择。我们开发了一种 knockoff 筛选程序，通过选择阈值 t_n 来控制错误发现率。选出的特征子集随后可用于采用传统聚类方法进行聚类识别，例如 k 均值或谱聚类算法。

接下来，我们在第4节探讨了 MM 检验统计量和特征筛选程序的理论性质。基于 MM 检验的筛选程序实现了选择一致性，这一点由定理1证明。定理2表明，我们的后选择聚类分析所得到的聚类结果，其汉明误差渐近收敛于零。此外，MM 检验的错误发现率得到了有效控制，达到预期水平，如定理3所示。

通过全面的模拟研究，我们证明，在第5节的各种情景中，MM检验在识别聚类相关特征方面具有更高的准确性。与基线方法相比，对筛选出的特征进行后续聚类分析，准确性和稳健性均显著提升。在第6节中，我们使用 34 个空间转录组数据集，对 MM检验与现有空间变异基因检测方法进行了基准比较。结果表明，MM检验始终优于现有方法。在第7节中，我们将 MM检验应用于三维小鼠脑数据集。我们发现，MM检验能够更准确地检测已知脑区的标记基因，因而有助于识别精细尺度三维脑结构。

2 数据描述与研究动机

本研究主要分析一个源自成年小鼠脑全脑分子图谱的数据集，该图谱采用空间转录组学技术生成，相关描述见Ortiz 等人（2020 年）。该数据集包含沿前后轴分布的 20 张冠状脑切片（图1A；详情见附录K）。每张切片均通过计算方法配准至 Allen 小鼠脑参考图谱（www.brain-map.org），从而获得所有测量组织位置的三维空间坐标。最终得到的三维数据集包含 11,061 个空间定位的组织位置上 11,879 个基因的表达谱，构成超高维数据结构。此外，测序基因表达为离散计数，而非连续数值。同时，大多数基因在任一给定位置的表达量为零或极低。平均而言，每个组织位置检测到 4,510 个基因，总计数为 14,739（图1B）。这反映出数据具有高度稀疏性且数值较小。总的来说，超高维、稀疏性以及计数数据数值较小等特点，使特征筛选——尤其是识别聚类信息基因——成为下游分析中的关键预处理步骤。

与传统二维空间转录组学实验相比，这个三维数据集以前所未有的方式展现了整个脑体积内的基因表达模式（图1C）。许多基因呈现空间局部化表达，局限于特定的解剖结构；这类基因通常称为空间变异基因（SVGs）。空间变异基因蕴含丰富的信息，可用于界定脑区和亚区（Zhang 等人，2023年），其中一些基因已被证实能够揭示此前未被识别的精细尺度分区（Qian 等人，2025 年）。由于许多解剖区域沿前后轴广泛延伸，单张二维切片不足以全面表征空间变异基因的完整空间分布。例如，*Ier5* 在三维空间中表现出强烈的空间

聚集于皮层区域（图 1C），但在二维切片中则没有此类模式（图 1D）。基于这些三维空间变异基因，无偏三维聚类能够重建完整的区域边界及相应的区域特异性基因表达模式。这类整合的空间转录组图谱为构建数字脑组织模型奠定了基础（Sharma 等人，2022年），其目标是共同呈现解剖与分子组织。这些数字图谱有望促进脑功能、疾病机制及治疗靶点发现方面的研究（Shi 等人，2023年）。

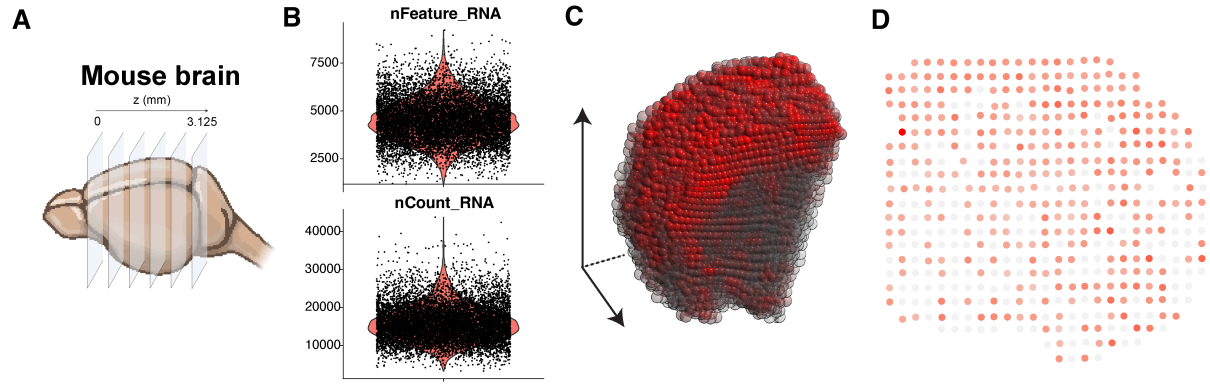

图 1：（A）来自 Ortiz 等人 (2020 年）的三维空间转录组数据集的实验设计。（B）各样本质量控制指标的小提琴图。上方面板显示 nFeature RNA（每个细胞检测到的基因数），下方面板显示 nCount RNA（UMI计数总数）。（C）*Ier5*的三维空间可视化；在该数据集中，该基因在同型皮层中富集。颜色强度表示表达水平，红色表示较高的表达水平。（D）*Ier5* 在二维切片中的表达可视化。

3 模型设置与 MM检验统计量

3.1 模型设置

令 $[I] = \{1, \ldots, I\}$ 表示与任意整数 I 关联的索引集。考虑 n 个相互独立的观测值 $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top \in \mathbb{R}^p$，对每个 $i \in [n]$，其中每个观测值都来自 K 个潜在簇之一。对于给定的簇索引 $k \in [K]$，定义属于簇 k 的样本索引集合为 $G_k = \{i \in [n] : k_i = k\}$，其中 $k_i \in [K]$ 表示

观测值 i 未知的聚类成员标签。令 $\alpha_k = \#_n^{G_k}$ 表示分配至簇 k 的样本的经验比例，其中 $\#G_k$ 表示 G_k 的基数。

令 F_{kj} 表示第 j 个簇中第 k 个特征的分布。若观测 i 属于簇 k，则对于所有 $j \in [p]$ 和 $i \in [n]$，x_{ij} 的边缘分布为 F_{kj}。令 μ_{kj} 和 τ_{kj} 分别表示 F_{kj} 的均值和方差。也就是说，若 $X_{kj} \sim F_{kj}$，则 $\mu_{kj} = \mathbb{E}(X_{kj})$ 且 $\tau_{kj} = \mathrm{Var}(X_{kj})$。对于给定的特征 j，当各簇中的均值 μ_{1j}、.....、μ_{Kj} 均相等时，该特征无助于区分不同簇，因此被视为与簇无关。据此，我们将与簇无关和与簇相关的特征集合定义如下：

$$S_0 = \{j : \mu_{1j} = \cdots = \mu_{Kj}\}, \quad S_1 = \{j : \mu_{kj} \neq \mu_{\ell j} \text{ for some } k, \ell \in [K]\}. \tag{1}$$

为识别相关特征，我们考虑以下一系列多重假设检验问题：

$$\mathbb{H}_{j0} : \mu_{1j} = \cdots = \mu_{Kj} \quad \text{vs.} \quad \mathbb{H}_{j1} : \mu_{kj} \neq \mu_{\ell j} \text{ for some } k, \ell. \tag{2}$$

我们将零假设 $\mathbb{H}_{j0}$ 下的模型称为齐性模型，将备择假设 $\mathbb{H}_{j1}$ 下的模型称为异质模型。由于检验是逐个特征进行的，为简化记号，我们省略下标 j，将 F_k、μ_k、τ_k、$\mathbb{H}_0$、$\mathbb{H}_1$ 分别记为 F_{kj}、μ_{kj}、τ_{kj}、$\mathbb{H}_{j0}$、$\mathbb{H}_{j1}$。

该检验问题（2）与经典假设检验的一个根本区别在于，聚类标签 k_i 不可观测。解决这一问题的一种策略是假设每个 F_k 都属于已知的参数族，从而能够通过基于似然的检验识别与聚类无关的特征。然而，如果真实分布与假设模型存在显著偏差，这些参数方法的表现可能会很差。作为一种灵活的替代方案，无分布假设方法，例如

准似然方法无需完整指定数据分布，即可实现稳健的参数估计和有效推断。因此，它们非常适用于复杂或异质的数据集，包括空间转录组学。

准似然方法假设方差 τ_k 是均值 μ_k的已知函数，具体为 $\tau_k = V(\mu_k; \phi)$，其中 $\phi > 0$ 是离散参数。方差函数$V(\mu; \phi)$通常具有以下形式： $\tau = V(\mu; \phi) = V_1(\mu) + \phi V_2(\mu)$，其中 $V_1(\mu) \geq 0$和$V_2(\mu) > 0$ 是已知的三阶连续可微函数。常见示例包括准泊松模型（$V_1(\mu) = 0$ 和 $V_2(\mu) = \mu$）以及准负二项模型（$V_1(\mu) = \mu$ 和 $V_2(\mu) = \mu^2$）。值得注意的是，准负二项模型的方差函数也对应于泊松混合模型的方差函数，后者广泛应用于组学数据分析（Sarkar 和 Stephens，2020 年）。指定的方差-均值函数导出分布族 $\mathcal{P}(\Theta, \Phi)$，其定义如下：

$$\mathcal{P}(\Theta, \Phi) = \{F\colon X \sim F, \mathrm{Var}(X) = V_1(\mathbb{E}(X)) + \phi V_2(\mathbb{E}(X)) \ \mathbb{E}(X) \in \Theta, \phi \in \Phi\}, \text{其中} (3)$$

$\Theta \subset \mathbb{R}$ 和 $\Phi \subset \mathbb{R}^+$ 是紧参数空间。接着，我们假设每个簇特异分布 F_k 都属于该族，即对于所有 $k \in [K]$， $F_k \in \mathcal{P}(\Theta, \Phi)$。

3.2 MM检验统计量

本节介绍如何使用 MM检验统计量评估假设框架（2）。从概念上讲，MM检验统计量通过方差函数 V（μ； ϕ）仅假定前两个矩之间存在函数关系，而非指定完整的似然，从而将似然比检验原理从参数推断扩展到半参数推断。因此，我们通过比较相互竞争的假设之间的准似然，而非传统似然，来构建检验统计量。我们首先建立了

准似然框架（Wedderburn，1974），该框架为推导 MM检验统计量提供了理论基础。

设 $\mu = \mathbb{E}(x)$、$\tau = \mathrm{Var}(x)$ 和 $\phi > 0$ 分别表示均值、方差和离散参数。随后我们将准似然函数定义为

$$Q(x;\mu,\phi) = \int_x^{\mu} \frac{x-t}{V(t;\phi)}\,\mathrm{d}t$$

并将准密度函数定义为 $f(x;\ \mu,\phi)\ = \exp(Q(x;\ \mu,\phi))$。请注意，$f(x;\ \mu,\phi)$ 并非真正的概率密度函数，因为其在 x 定义域上的积分通常不等于 1。这种准似然构造保留了基本的对数似然性质，即使底层概率模型未完全指定，也能便于进行推断。

设 $f(x;\ \mu_k,\phi_k)$ 表示第 k 个分量（$k \in [K]$）的准密度函数。随后我们将异质模型的总体准密度函数定义为

$$\varphi(x;\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\phi}) = \sum_{k=1}^{K} \alpha_k f(x;\mu_k,\phi_k),$$

其中 $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K)$ $\boldsymbol{\xi} = (\mu_1,\ldots,\mu_K)$，且 $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_K)$。请注意，$\varphi(x;\ \boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\phi})$ 与标准混合模型的密度函数具有相似形式。该异质模型的准似然函数随后定义为

$$l_n(\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\phi}) = \sum_{i=1}^{n} \log\ \varphi(x_i;\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\phi}).$$

在齐性模型 $\mathbb{H}_0$ 下，其中 $\mu_1 = \cdots = \mu_K = \mu_0$，准似然函数变为 $l_n(\boldsymbol{\alpha},\boldsymbol{\xi}_0,\boldsymbol{\phi})$，其中 $\boldsymbol{\xi}_0 = (\mu_0,\ldots,\mu_0)$。若所有离散参数 ϕ_k 均等于一个共同的 ϕ，我们将记号 $l_n(\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\phi})$ 简化为 $l_n(\boldsymbol{\alpha},\boldsymbol{\xi},\phi)$

类似于经典似然比检验，我们将准似然比（QLR）检验统计量定义为 $\mathrm{QLR} = -2(L_0 - L_1)$，其中 L_0 和 L_1 分别是原假设（$= -2$）和备择假设（$L_0 - L_1$）下的最大准对数似然

值。由于准似然函数形式特殊，直接最大化准对数似然通常难以实现。为克服这一挑战，我们采用 MM算法这一强大的优化技术，在每一步通过最大化代理函数来迭代改进参数估计值（附录A.2）。

应用MM算法的一个挑战在于，$\boldsymbol{\phi}$ 的更新步骤缺乏闭式解，且计算量很大。为解决这一问题，我们通过ˆ约束 $\phi_1 = \cdots = \phi_K = \phi$来简化优化，并在整个更新过程中将 ϕ 固定为工作离散参数 $\hat{\phi}$ˆ。工作离散参数 ϕ 通过整合空间坐标等辅助信息来构建（第3.3节）。经过这一简化，MM更新变为：

$$w_{ki}^{(t+1)} = \frac{\alpha_k^{(t)} f(x_i; \mu_k^{(t)}, \hat{\phi})}{\varphi(x_i; \boldsymbol{\alpha}^{(t)}, \boldsymbol{\xi}^{(t)}, \hat{\phi})}, \alpha_k^{(t+1)} = n^{-1} \sum_{i=1}^{n} w_{ki}^{(t+1)}, \mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ki}^{(t+1)} x_i}{\sum_{i=1}^{n} w_{ki}^{(t+1)}}. \tag{4}$$

ˆ显然，在工作离散参数 ϕ下，MM更新无需额外优化即可轻松计算，因而计算效率很高。此外，我们证ˆ明，基于工作离散参数 ϕ定义的MM检验，对特征选择及下游聚类具有强有力的理论保证（第4节）。

经过 $T_n > 0$ 次MM更新后，我们将 $\mathbb{H}_1$ 下最大化的准对数似然记为$L_1 = l_n\left(\boldsymbol{\alpha}^{(T_n)}\right.$ $\left.\boldsymbol{\xi}^{(T_n)}, \hat{\phi}\right)$。令ˆ $\boldsymbol{\xi}_0 =$(¯x.....， ¯x)为 $\boldsymbol{\xi}_0$的估计量，并令 $l_n\left(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\xi}}_0, \hat{\phi}\right)$表示 $\mathbb{H}_0$下的准对数似然 L_0。于是，MM检验统计量定义为

$$\mathrm{MM}_n^{(T_n)} = 2\left\{l_n\left(\boldsymbol{\alpha}^{(T_n)}, \boldsymbol{\xi}^{(T_n)}, \hat{\phi}\right) - l_n\left(\boldsymbol{\alpha}_0, \hat{\boldsymbol{\xi}}_0, \hat{\phi}\right)\right\}.$$

利用 MM 检验统计量，我们提出以下聚类特征筛选程序。令 $\mathrm{MM}^{(T_n)}{}_{nj}$ 为第 j个特征的 MM 检验统计量，对应于

假设检验问题 (2)。给定阈值 $t_n > 0$，我们定义估计的聚类相关特征集合 $\hat{S}_1(t_n) = \left\{1 \le j \le p : \mathrm{MM}_{nj}^{(T_n)} \ge t_n\right\})$，它是 S_1的一个估计量。如果某个特征 j 的 MM 检验统计量 $\mathrm{MM}^{(T_n)}{}_{nj} \ge t_n$，则将其纳入 $\hat{S}_1(t_n)$；否则，该特征将被筛除。

3.3 利用空间信息设定工作离散参数 ϕ

单细胞组学与空间组学的一个关键区别在于，后者纳入了空间上下文，通常通过位置图表示，并常辅以组织图像；这些信息提供了分子表达谱以外的结构信息。我们可以系统地利用这些空间信息，提升簇相关基因识别的效果。我们通过辅助距离矩阵 $\mathbf{D} = [D_{ij}] \in \mathbb{R}^{n\times n}$ 将空间信息形式化，其中 $D_{ii} = 0$ 对所有 i 均成立，且 $D_{ij} > 0$ 对所有 $i \ne j$ 均成立。该矩阵 $\mathbf{D}$ 编码观测值之间的成对辅助距离，并假设其包含关于潜在簇结构的微弱但有用的信号，从而弥补单纯基于表达的方法在特征选择方面的不足。$\mathbf{D}$ 矩阵的构建兼具灵活性并依赖于数据，可适配能够捕捉聚类信号的多种信息来源，包括空间坐标（二维或三维）、组织学图像特征及其他模态特异性测量。

示例 1（空间转录组数据）。 在空间转录组学中，通过对空间点处的高维表达向量（x_{i1}、…、x_{ip}）进行聚类来识别空间域，每个空间点的坐标为（u_i, v_i）（二维）或（u_i, v_i, w_i）（三维）。空间点之间一种自然的辅助距离是欧几里得距离，其定义为：$D_{ij}^S = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$对于二维空间数据，或 $D_{ij}^S = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2 + (w_i - w_j)^2}$ 对于三维空间数据。

此外，许多空间转录组数据集还包含配对的组织学图像，例如 *H&E*染色图像。我们可以使用现有的神经网络为每个空间点 i提取局部图像表示 $\boldsymbol{z}_i \in \mathbb{R}^d$，并将辅助距离 D_{ij}^I 定义为 $\boldsymbol{z}_i$ 与 $\boldsymbol{z}_j$之间的欧几里得距离： $D_{ij}^I = \|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2$. 另一种方法是结合空间坐标和组织学图像，并将距离定义为 $D_{ij} = D_{ij}^S + D_{ij}^I$.

我们利用辅助距离矩阵 **D**来获取工作离散参数 $\hat{\phi}$。在介绍估计过程之前，先阐述其理论基础。我们将合并均值、方差和离散度定义为

$$\mu_0 = \sum_{k=1}^{K} \alpha_k \mu_k, \quad \tau_0 = \sum_{k=1}^{K} \alpha_k \tau_k + \sum_{k=1}^{K} \alpha_k (\mu_k - \mu_0)^2, \quad \phi_0 = \frac{\tau_0 - V_1(\mu_0)}{V_2(\mu_0)}, \tag{5}$$

并令 $\boldsymbol{\xi}_0 = (\mu_0, \ldots, \mu_0)$。假设所有簇具有相同的离散参数，即 $\phi_1 = \cdots = \phi_K = \phi$，则准泊松模型意味着量 $\phi = \phi_0 - \frac{\sum_{k=1}^{K} \alpha_k (\mu_k - \mu_0)^2}{V_2(\mu_0)}$，这表明离散参数 ϕ 可估计为合并离散度 ϕ_0 与按方差缩放的均值差平方和之差，其中， μ_k 表示簇特异性均值， μ_0 表示合并均值， α_k 表示簇比例，而 $V_2(\mu_0)$ 表示方差函数。

为近似这一关系，我们采用辅助距离矩阵**D** 来估计局部均值与全局均值之间的平方偏差。具体而言，对于每个样本 i，我们将其邻域 $\mathfrak{N}(i)$定义为其 $r_n = n^\beta$ 个最近邻的集合，其中 $0 < \beta < 1$和 $\beta = 0.9$ 是默认值，由距离矩阵**D**确定。合并离散度 ϕ_0 可估计为 $\hat{\phi}_0 = (\hat{\tau}_0 - V_1(\bar{\mathbf{x}}))/V_2(\bar{\mathbf{x}})$，其中 $\hat{\tau}_0 = n^{-1} \sum_{i=1}^{n} (x_i - \bar{\mathbf{x}})^2$ 是样本方差。因此，离散参数可估计为

$$\hat{\phi} = \max \left\{ \hat{\phi}_0 - \sqrt{r_n} \cdot \frac{1}{n} \sum_{i=1}^{n} (\bar{\mathbf{x}} - \hat{\mu}_i)^2 \Big/ V_2(\bar{\mathbf{x}}),\, m_\phi \right\}, \tag{6}$$

其中 $\hat{}\mu_i = $ 均值$\{x_\ell : \ \ell \in \mathfrak{N}(i)\}$, $\sqrt{r_n}$ 是一个用于提高检测功效的缩放因子，$\hat{}$ 而 $m_\phi > 0$ 是一个确保 ϕ非负性的常数。

直观地说，如果辅助数据 **D** 中所包含的簇信息非常强，那么大多数邻域 $\mathfrak{N}(i)$'s 中将只包含来自单个簇的样本，并且 $\hat{}\mu_i$ 将接近某个 μ_k's。因此， $n^{-1}\sum_{i=1}^{n}(\bar{}\mathbf{x} - \hat{}\mu_i)^2$ 接近 $\sum_{k=1}^{K}\alpha_k(\mu_0 - \mu_k)^2$。一般而言，**D** 中的簇信息较弱， $n^{-1}\sum_{i=1}^{n}(\bar{}\mathbf{x} - \hat{}\mu_i)^2$ 往往会显著低估$\sum_{k=1}^{K}\alpha_k(\mu_k - \mu_0)^2,$，因此我们加入缩放因子 $\sqrt{r_n}$，以放大 $\mathbb{H}_1$下的信号。

—

或者，在零假设 $\mathbb{H}_0$下，该特征与簇无关：所有簇具有相同的均值，且合并离散度 ϕ_0 等于离散参数 ϕ。在这种情况下，$\hat{}$ 局部均值 $\hat{}\mu_i$ 也往往接近合并均值 μ_0，并且 ϕ_0 接近 ϕ_0。因此，即使使用缩放因子 $\sqrt{r_n}$，经验量 $\frac{\sqrt{r_n}}{n}\sum_{i=1}^{n}(\bar{}\mathbf{x} - \hat{}\mu_i)^2$ 也接近 0，$\hat{}\hat{}$ 而工作离散参数 ϕ 接近离散参数 ϕ。

— —

因此，在零假设和备择假设下，工作离散参数 ϕ 都可作为真实离散参数的合理替代。理论分析表明，采用这一工作离散参数，MM检验能够实现特征选择一致性，并取得较高的聚类准确率。

3.4 通过 knockoff 控制错误发现率

全面的特征选择框架不仅需要识别聚类相关特征，还需要为其赋予有效的 p值，并且至关重要的是，控制错误发现率。回顾一下， $S_0 \subset [p]$ 表示真实零假设特征的集合（即与簇无关的特征集合），$\hat{}\hat{}$ 而 S_1 是聚类相关特征的估计集合。令 $V = S_0 \cap S_1$ 为错误拒绝特征的集合。错误发现率正式定义为所有选择结果中错误发现所占比例的期望：错误发现率 $= \mathbb{E}\left(\frac{\#V}{\max\{\#\hat{}S_1,\ 1\}}\right)$.

———

通常，由于对数似然函数具有非凸性，MM 检验统计量的 p值难以计算。为解决这一推断难题，我们开发了*MM* 检验 *knockoff* 程序，这是对 pCandes 等人提出的无需值的 knockoff 框架(2018 年)的扩展。该程序无需显式计算 p 值或使用渐近近似，即可严格控制错误发现率。令$\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p) = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times p}$。MM 检验程序包含以下步骤:

1. **Knockoff 构造**。 对于每个特征 $j \in [p]$，我们生成 knockoff 特征$\widetilde{\mathbf{X}}_j \in \mathbb{R}^n$ ，方法是从 $\mathbf{X}_j$进行有放回抽样。由此得到 knockoff 矩阵$\widetilde{\mathbf{X}} = (\widetilde{\mathbf{X}}_1, \ldots, \widetilde{\mathbf{X}}_p) \in \mathbb{R}^{n \times p}$。

2. **MM 检验统计量计算**。 我们使用增广矩阵 $[\mathbf{X} \ \widetilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$计算 MM检验统计量，得到 $\left(\mathrm{MM}^{(T_n)}_{n1}, \ldots, \mathrm{MM}^{(T_n)}_{np}, \mathrm{MM}^{(T_n)\star}_{n1}, \ldots, \mathrm{MM}^{(T_n)\star}_{np}\right)$,并定义经缩放的 knockoff 统计量，以增强错误发现率控制:

$$\left(\widetilde{\mathrm{MM}}^{(T_n)}_{n1}, \ldots, \widetilde{\mathrm{MM}}^{(T_n)}_{np}\right) = \left(1 + \frac{1}{\log n}\right)\left(\mathrm{MM}^{(T_n)\star}_{n1}, \ldots, \mathrm{MM}^{(T_n)\star}_{np}\right).$$

3. **阈值确定**。 对于目标错误发现率水平 $q_0 \in (0,1)$，我们定义阈值 H_{q_0}:

$$H_{q_0} = \min\left\{h \in \mathcal{M} : \frac{\#\{j : \widetilde{\mathrm{MM}}^{(T_n)}_{nj} \geq h\}}{\#\{j : \mathrm{MM}^{(T_n)}_{nj} \geq h\} \vee 1} \leq q_0\right\}, \tag{7}$$

其中 $\mathcal{M} = \{|\mathrm{MM}^{(T_n)}_{nj}| : \ j \in [p]\}$ 是唯一非零绝对检验统计量的集合，且 $H_{q_0} = +\infty$ 表示没有阈值满足该准则。

我们的框架采用简化的 knockoff 生成方法，通过直接重采样避免了显式联合分布建模的计算复杂性（Candes 等人，2018 年）。如第 4 节所示，这种计算高效的程

序在提升实际实现效率的同时，仍能为错误发现率控制提供理论保证。错误发现率控制的特征选择集使用阈值H_{q_0}，$\hat{S}_1(H_{q_0}) = \left\{ j \in [p] : \mathrm{MM}_{nj}^{(T_n)} \geq H_{q_0} \right\}$.）进行形式化定义。该数据自适应阈值 H_{q_0} 提供了经过实证校准的错误发现率控制，因此尤其适用于底层特征分布未知的实际应用。

4 MM检验的理论性质

设 $\{a_n\}$ 和 $\{b_n\}$ 为正数列。若 $a_n/b_n \to 0$，则记为 $a_n \ll b_n$；若存在某个常数 $C > 0$使得$a_n \geq C|b_n|$，则记为 $a_n \gtrsim b_n$；类似地，若 $a_n \leq C|b_n|$，则记为 $a_n \lesssim b_n$；若$a_n \gtrsim b_n$ 和 $a_n \lesssim b_n$ 均成立，则记为 $a_n \asymp b_n$。

在推导 MM检验特征筛选程序的选择一致性保证之前，我们先介绍四个条件（条件1-4），这些条件共同确保渐近性质。完整的数学表述和详细的正则条件见附录B。

条件1是矩约束条件；它假设每个特征的分 L_q 布的范数均由常数 M_q一致有界，从而确保尾部按多项式速度衰减。[1] 条件2对空间信息和簇间差异提出了较弱的要求。基本上，它假设邻域 $\mathfrak{N}(i)$中的局部均值与全局均值 μ_0 之间的差异平均而言超过阈值 $r_n^{-\vartheta}$，其中r_n 是邻域大小，$\vartheta < 1/2$ 是一个常数。这意味着空间数据$\mathbf{D}$ 必须包含最低程度的簇信息；否则，$\mathfrak{N}(i$中的大多数局部均值将非常接近全局均值 μ_0。

条件3是簇可分性条件，假设分离信号 η_n

[1]如果假设分布具有次指数尾部，则定理1中的概率界可改进为呈指数衰减速率。

在 K 簇之间足够强。条件4是零齐次性条件，对于与聚类无关的特征，要求其在不同簇中的均值和方差均相同，以确保 knockoff方法能够有效控制错误发现率。

现在，我们为特征筛选流程建立理论保证，并推导出严格的渐近性质，以确保模型选择一致性。

定理1. 假设条件*1–2*成立。设 $0 < \varpi < 1/2 - \vartheta$ 为一个常数。当n 足够大时，对于 $r_n^{\vartheta}(\log n)^2 < T_n < r_n^{1/2-\varpi}\log n$ 以及 $t_n = r_n^{-\kappa}$, $-1 + 3\vartheta < \kappa < 5/2$，有 $\mathbb{P}\left(S_1 \neq \hat{} S_1(t_n)\right) \lesssim (p-s)\, r_n^{-c_1 q + \beta^{-1}} + s r_n^{-c_2 q + \beta^{-1}} + o(1)$，其中 $r_n = n^{\beta}$, $s = \#S_1$和 c_1, c_2 是两个正的常数，仅取决于 κ 和 ϑ。

^定理1 给出了 $S_1 \neq S_1(t_n)$的概率，其中 S_1 表示^聚类相关特征的集合，而 $S_1(t_n)$ 表示我们的方法在空间情形下选出的特征。如果 $p \asymp n^{\omega}$ 且 $0 < \omega + 1 < q\beta(c_1 \wedge c_2)$, 则有 $\mathbb{P}\left(S_1 = \hat{} S_1(t_n)\right) \to 1$，因此$n \to \infty$，换言之，我们可以实现模型选择一致性。

接下来，我们推导将 MM 检验选定的特征子集用于聚类时，聚类准确率的理论保证。令 $\hat{}\mathbf{k}^{\mathrm{MM}}$ 表示在 MM 检验选定特征的 PCA 降维子空间中进行 K 均值聚类后得到的估计聚类分配向量（Jin and Wang，2016年）。参数 η_n 量化了相对于簇内变异的最小簇间分离度（正式定义见附录A），作为决定分类准确率的聚类信号强度。为量化聚类准确率，我们采用估计聚类分配与真实聚类分配之间的汉明距离，$\mathrm{Hamm}^*\left(\hat{}\mathbf{k}^{\mathrm{MM}}, \mathbf{k}\right) = \min_{\pi}\left\{\sum_{i=1}^{n}\mathbb{P}\left(\hat{}(k_i)\right)\right\}$，其中 π 是 中的任意排列 $\{1, \ldots, K\}$。

定理2。 假设 $s = n^{\gamma}$ 和簇分离信号 $\eta_n \gtrsim s^{1/2} n^{-\vartheta}$, $0 \leq \vartheta < 1/4$.对于任意 $\gamma > 4\vartheta$ 和 $q > (\epsilon\gamma)^{-1} \vee (\gamma/\epsilon)$，其中 $\epsilon = 1/4$–最大值$\{\vartheta, \vartheta/\gamma\}$，当 n 充分

大时，在条件*1–3*下，我们有：$n^{-1}\,\mathrm{Hamm}^*\left(\hat{\mathbf{k}}^{\mathrm{MM}},\ \mathbf{k}\right) \lesssim \mathbb{P}\left\{\hat{S}_1(t_n) \neq S_1\right\} + o(1)$。

定理 2 指出，使用筛选出的特征进行聚类时，所得聚类分配与真实聚类分配之间的汉明距离，上界为模型误选的概率加上聚类过程本身引入的可忽略误差项。第二项对应于使用所有聚类相关特征得到的预言机聚类误差。因此，基于所提方法得到的聚类结果几乎与使用全部相关特征时一样准确。定理 2 的推导面临重大技术挑战，因为我们的框架适用于仅满足多项式矩条件的重尾分布，并大幅推广了以往依赖高维统计学中常见的次高斯或次指数假设的结果（Jin and Wang，2016年）。完整的数学推导及相关技术引理见附录G，其中我们建立了有限样本保证和渐近性质。

现在，我们为 MM 检验程序下的错误发现率控制建立严格的理论保证。

定理 3. 假设条件*1–4*成立，对于任意 $q_0 > 0$，当 $s \to \infty$，选择阈值 H_{q_0}，其定义见 (7) 可确保 错误发现率 $\leq q_0 + o(1)$。

定理 3 指出，随着样本量 q_0 $n \to \infty$，错误发现率（FDR）渐近受控于目标水平。证明依赖于建立原始统计量与 knockoff 统计量之间的渐近分布收敛。我们的证明策略解决了基于自助法的 knockoff 生成所带来的重大技术挑战：这种生成方式只能保证特征与其 knockoff 之间渐近而非精确的可交换性，并且需要新的集中不等式和离散化技术。完整证明见附录H。

5 模拟

本节通过模拟空间转录组学数据，评估 MM检验的性能。我们将其与一系列广泛使用的方法进行了比较分析（Chen 等人，2024年），包括用于空间变异基因检测的 Moran（Miller 等人，2021 年）、SPARK-X（Zhu 等人，2021 年）、Binspect（Dries 等人，2021 年）和 SOMDE（Hao 等人，2021 年），以及用于特征筛选的 SCFS（Liu 等，2022年）。对于 MM检验统计量，我们采用默认的方差函数 $V(\mu) = \mu + \phi\mu^2$。为评估 MM检验对方差设定错误的稳健性，我们考虑了另一种设定 $V(\mu) = \phi\mu$。稳健性分析的完整模拟细节以及各方法的计算时间比较分别见附录 J.1 和 J.2。附录 J.3 介绍了一项三维模拟，旨在进一步考察 MM检验统计量的性能。

在模拟中，我们考虑了两种不同的空间布局（图 S1）。布局1是一个方形区域，由 $n = 900$ 个空间点组成，并被划分为 $K = 5$个等大的矩形区域。布局 2 来源于小鼠脑矢状切片的带注释的空间转录组数据，由 $n = 2{,}687$ 个空间点组成，分布在$K = 8$ 个解剖学定义区域中。在所有模拟中，我们考虑了三种维度设置，即 $p = 3{,}000$、10, 000 或 $30,000$。与簇相关/无关的特征由负二项（NB）分布或泊松对数正态（PLN）分布生成。我们随机将 30% 的元素设为零，以提高模拟中的稀疏性。我们考虑了两种信号强度设置，分别量化低信号和高信号强度条件下的组间差异。在每种模拟设置中，我们生成了 100 个数据集。详细的模拟设置见附录 I。由于计算负担较高，Binspect 和 Moran 被排除在评估之外

布局 2 配置，数量为$p = 10{,}000$ 或 $p = 30{,}000$。

5.1 特征筛选性能

我们通过精确率-召回率曲线下面积（AUPRC）分析评估了特征筛选性能。特征按其各自的检验统计量或p值排序，并通过改变这些排序阈值生成精确率-召回率曲线。MM检验在所有模拟情景中均表现出更优或相当的性能，其优势在更具挑战性的条件下尤为明显：复杂的空间布局（布局 2）、较低的信号强度以及较高的特征维度（表 S1）。值得注意的是，在最具挑战性的情景（布局 2、低信号强度、$p = 30{,}000$ 个特征）中，MM检验仍保持稳健性能，均值 AUPRC 达到 0.75。紧随 MM检验之后，SPARK-X 表现出次优性能（0.42），在大多数情景中均较为有效。

接下来，表1 总结了每种方法的统计功效（正确保留的特征数与 s之比）以及错误发现率。对于 MM检验，如果特征通过 knockoff方法被选中，则予以保留，条件为 $q_0 = 0.05$。对于 SCFS，使用其默认参数选择聚类相关特征。对于 SPARK-X、Binspect、SOMDE 和 Moran，如果经 Benjamini–Hochberg 校正后的 p值小于 0。05，则保留这些特征。与其他方法相比，MM检验通常能够保留更多真正的聚类相关特征，同时仅产生极少的假阳性。SPARK-X 和 SCFS 往往较为保守，假阳性较少，但经常会漏掉许多真正的聚类相关特征，尤其是在更具挑战性的设置下。空间布局的复杂程度显著影响了 Moran 和 Binspect 的表现。这些方法在较简单的布局1模拟中表现尚可，但在更复杂的布局 2情景下则表现不佳。

表1：不同方法在100次重复中的功效均值和错误发现率。目标错误发现率水平为0.05。括号中的数字是100次重复中功效和错误发现率的标准差。表中的数值为实际值 $\times$ 100。由于这些情形的计算成本过高，Moran和Binspect未在布局2模拟中针对 $p = 10{,}000$和 $p = 30{,}000$进行评估。

p		SCFS	SPARK-X	Moran	Binspect	SOMDE	MM检验
布局-1，高信号							
3,000	功效	99 (4.5)	100 (0.5)	100 (0.1)	96 (1.6)	16 (1.9)	100 (0.0)
	FDR	0 (0.0)	0 (0.0)	30 (2.7)	41 (3.2)	83 (1.9)	5 (2.3)
10,000	功效	98 (3.3)	100 (0.6)	100 (0.2)	98 (1.0)	15 (1.8)	100 (0.0)
	FDR	0 (0.0)	0 (0.0)	45 (2.3)	61 (2.2)	83 (1.9)	5 (1.8)
30,000	功效	69 (5.2)	98 (1.1)	100 (0.2)	99 (0.8)	6 (2.2)	100 (0.0)
	FDR	0 (0.0)	0 (0.0)	62 (1.9)	77 (1.1)	92 (2.8)	5 (2.1)
布局-1，低信号							
3,000	功效	0 (0.3)	55 (3.3)	49 (4.6)	27 (3.9)	0 (0.1)	90 (3.4)
	FDR	0 (0.0)	0 (0.0)	38 (3.9)	61 (4.3)	10 (30.3)	6 (2.8)
10,000	功效	0 (0.0)	42 (3.5)	40 (4.2)	30 (3.9)	0 (0.0)	69 (5.7)
	FDR	0 (0.0)	0 (0.0)	58 (3.0)	79 (2.2)	0 (0.0)	6 (2.7)
30,000	功效	0 (0.0)	33 (2.9)	35 (4.5)	32 (3.4)	0 (0.0)	52（6.9）
	FDR	0（0.0）	0（0.0）	76（3.0）	90（1.2）	2（14.1）	5（3.0）
布局-2，高信号							
3,000	功效	20（0.7）	69（1.5）	22（1.6）	6 (1.2)	43 (1.5)	100 (0.1)
	FDR	0 (0.0)	0 (0.0)	0 (0.2)	0 (0.0)	9 (2.8)	2 (1.0)
10,000	功效	15 (1.4)	62 (1.6)	NA	NA	8 (0.9)	98 (0.7)
	FDR	0 (0.0)	0 (0.0)	NA	NA	83 (1.9)	3 (1.0)
30,000	功效	0 (0.0)	54 (1.5)	NA	NA	6 (0.9)	90 (1.7)
	FDR	0 (0.0)	0 (0.0)	NA	NA	83 (2.3)	4 (1.6)
布局-2，低信号							
3,000	功效	0 (0.1)	36（1.9）	0（0.0）	0（0.2）	0（0.1）	87（10.0）
	FDR	0（0.0）	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	5 (2.2)
10,000	功效	0 (0.0)	28 (1.7)	NA	NA	0 (0.0)	66 (2.5)
	FDR	0 (0.0)	0 (0.0)	NA	NA	0 (0.0)	6 (2.4)
30,000	功效	0 (0.0)	23 (1.6)	NA	NA	0 (0.0)	57（6.9）
	FDR	0（0.0）	0（0.0）	NA	NA	2（14.1）	4（1.8）

模拟进一步表明，MM检验能够在名义水平 0.05 下有效控制错误发现率，从而验证了我们的统计理论，并提供了一种数据驱动的阈值确定方法。

5.2 特征筛选可改进聚类分析

通过错误发现率控制进行特征筛选后，我们进一步在由缩减后的特征集导出的十个主成分上进行了 k均值聚类。簇的数量固定为与真实簇数量一致。表2 展示了各方法得到的簇与真实分组之间的调整兰德指数（ARI）值。MM检验始终

优于所有场景中的其他方法，在低信号、高维或空间复杂配置下尤其出色。无筛选方法在较低维度下表现尚可，但在高维环境中的性能显著下降，凸显了聚类前严格进行特征筛选的重要性。SPARK-X表现稳定，但在具有挑战性的模拟条件下，其准确性显著低于MM检验，反映出它在识别细微簇边界方面的局限性。SCFS在大多数场景中表现不佳，主要是因为其过于保守的特征保留标准，导致真实的簇相关特征被不成比例地排除。

表 2：不同方法在 100 次重复实验中的调整兰德指数均值。括号中的数字为 100 次重复实验中调整兰德指数的标准差。表中的数值为实际值 $\times$ 100。由于计算成本过高，在布局 2 模拟中，Moran 和 Binspect 未在 $p = 10{,}000$ 和 $p = 30{,}000$ 的情形下进行评估。

p	不进行筛选	Oracle	SCFS	SPARK-X	Moran	Binspect	SOMDE	MM检验
布局-1，高信号								
3,000	95 (5.3)	100 (0.2)	99 (3.4)	**100 (0.2)**	**100 (0.3)**	**100 (0.4)**	53 (4.0)	**100 (0.2)**
10,000	81 (4.2)	100(0.2)	**100 (1.5)**	**100 (0.2)**	**100 (0.3)**	**100 (0.3)**	51 (4.9)	**100 (0.2)**
30,000	53 (1.9)	100 (0.3)	96 (4.4)	**100（0.3）**	**100（0.3）**	99 (0.4)	23 (8.7)	**100 (0.3)**
布局-1，低信号								
3,000	46 (2.1)	85 (1.9)	0 (0.8)	62 (5.0)	61 (4.2)	36 (5.5)	0 (0.0)	**84 (2.3)**
10,000	4 (1.9)	85 (2.4)	0 (0.0)	48 (5.0)	51 (5.0)	34 (5.1)	0 (0.0)	**75 (3.7)**
30,000	1 (0.4)	85 (2.1)	0 (0.0)	40 (3.6)	41 (6.0)	24 (6.4)	0（0.0）	**64（6.5）**
布局-2，高信号								
3,000	83（9.0）	97（3.9）	47 (8.5)	85 (7.0)	53 (6.4)	31 (6.1)	36 (5.0)	**98 (3.4)**
10,000	49 (4.3)	97 (4.1)	41 (7.6)	82 (6.2)	NA	NA	33 (4.1)	**96 (5.0)**
30,000	4 (0.9)	96 (5.1)	0 (0.0)	78 (6.7)	NA	NA	11 (4.5)	**96 (5.3)**
布局-2，低信号								
3,000	42 (3.8)	92 (1.6)	0 (0.2)	48 (7.0)	0 (0.0)	0 (0.7)	7 (2.1)	**85 (8.3)**
10,000	6 (1.3)	91 (3.2)	0 (0.0)	40 (6.6)	NA	NA	7 (2.5)	**73 (6.1)**
30,000	0 (0.1)	91 (3.6)	0 (0.0)	34 (6.7)	NA	NA	2 (1.4)	**65 (8.0)**

6 真实空间转录组数据基准测试

本节展示了在真实空间转录组数据集上对空间变异基因识别方法进行基准测试的结果，并从检测准确性、假阳性控制和空间聚类能力等方面评估其表现。根据基准测试

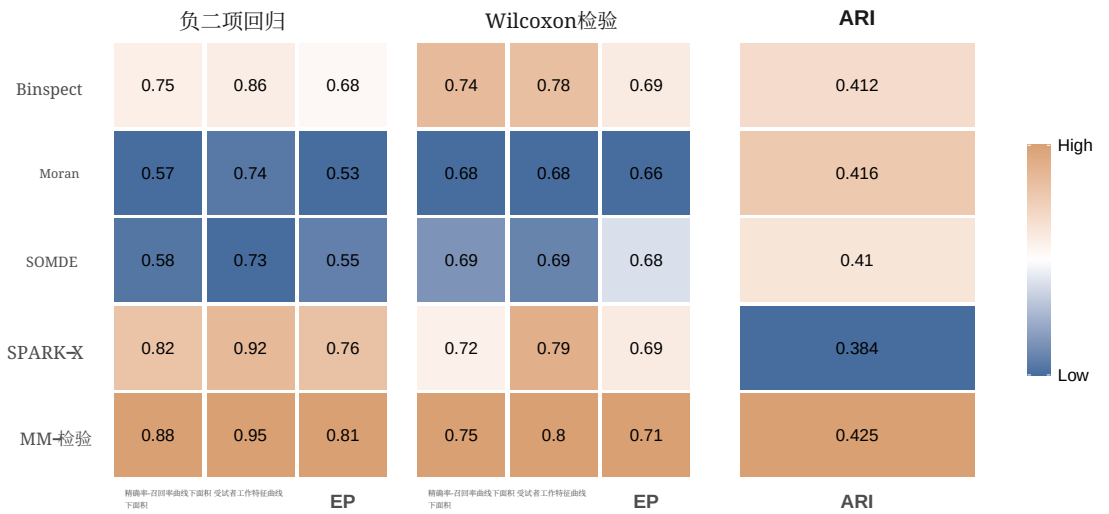

图 2：在 34 个真实空间转录组数据集上对空间变异基因识别方法进行基准测试的结果。纵轴表示各方法的精确率-召回率曲线下面积、受试者工作特征曲线下面积和早期精确率（EP）的均值，这些指标分别针对每种银标准计算（Wilcoxon检验和负二项回归）。最后一个面板展示了基于各方法识别出的空间变异基因进行空间聚类所得的调整兰德指数均值。数值越高表示性能越好。

由Chen 等人(2025 年)制定的方案，我们在 34 个公开可用的数据集上评估了这些方法。这些数据集均包含空间域的专家注释，涵盖了多种实验方案、组织类型、样本量和空间分辨率。数据集包括人脑数据集（背外侧前额叶皮层）（Maynard等人，2021 年）、发育中小鼠胚胎数据集（DME）（Srivatsan 等人，2021 年）以及成年小鼠脑（MAM）（Ortiz 等人，2020 年）。

为评估空间变异基因检测准确度，我们采用了两种空间域标记基因的“银标准”，如Chen 等人(2025 年所述。其中一种银标准通过使用 Wilcoxon 秩和检验，对每个已注释空间域与所有其他空间域进行差异表达分析生成。另一种则通过对基因表达进行空间域负二项回归构建。在这两种方法中，均选取 Benjamini–Hochberg 校正后的 p 值低于 0.05 的基因作为银标准基因。

我们采用了来自Chen 等人（2025 年）的三项评估指标：精确率-召回率曲线下面积（AUPRC）、受试者工作特征曲线下面积

（AUROC），以及早期精确率（EP）。需要注意的是，早期精确率衡量的是在前K个已识别空间变异基因中真阳性所占的比例，其中 K 对应于每个银标准所识别的空间变异基因数量。这些评估指标的数值越高，表明性能越好。

图 2展示了各方法在34个数据集上的精确率-召回率曲线下面积、受试者工作特征曲线下面积和早期精确率均值，并针对每种银标准分别计算。MM检验始终优于其他方法。按照 Wilcoxon检验和负二项回归标准，MM检验均取得最高的精确率-召回率曲线下面积、受试者工作特征曲线下面积和早期精确率，凸显了其在空间变异基因检测中的稳健性。为评估空间变异基因检测中的假阳性控制，我们使用置换数据集进行了零假设分析。如附录K所述，MM检验、SPARK-X 和 SOMDE 能有效控制假阳性，而 Moran 和 Binspect 则显示出偏高的假发现率，这与它们在模拟研究中的表现一致（图 S3）。

最后，我们评估了各方法基于识别出的空间变异基因促进空间聚类的效果。按照 Chen 等人（2025 年）的方案，我们使用 Seurat流程进行空间聚类，并采用 Louvain 算法进行优化（Butler等人，2018 年），所依据的是各方法识别出的排名前 2,000 个空间变异基因。通过推断出的簇与专家标注的空间域之间的调整兰德指数来量化聚类准确率。为确保比较公平，我们将 Louvain算法的分辨率参数从 0.5 调整至 1.5，并为每种方法选取使调整兰德指数达到最大的参数值。如图2所示，所提出的 MM检验取得了最高的调整兰德指数，表明它能够有效保留对空间域识别至关重要的空间信息基因。

7 MM检验可准确检测精细的三维空间域

我们将 MM检验应用于成年小鼠脑数据集，该数据集包含 20 张冠状切片，详见第2节。MM检验的辅助距离矩阵选用空间位置之间的三维欧几里得距离。使用 knockoff 程序进行基因选择后，在错误发现率水平 0.05 下，我们采用 Louvain算法对空间位置进行聚类，分辨率参数设为 1.4。该过程识别出 20 个空间聚类（图3A）。作为比较，我们调整了现有的空间基因检测方法——SPARK-X、Moran 和 Binspect，使其能够通过纳入空间点的三维空间坐标处理三维数据。此外，我们还纳入了两种非空间特征选择方法：SCFS 和高变基因（HVG）方法（Butler等人，2018 年）。由于难以将 SOMDE 调整为适用于三维空间数据，因此未将其纳入比较。对于除 HVG 和 SCFS 外的所有竞争方法，我们选取 Benjamini–Hochberg 校正后 p值低于 0.05 的基因。对于 SCFS，使用默认阈值 0.9 来识别空间变异基因。对于 HVG，选取变异性最高的前 2,000 个基因。我们也对每种竞争方法应用 Louvain算法，并调整分辨率参数，使每种方法均识别出 20 个簇（图 S4）。

总体而言，MM检验生成的簇与已知脑区更为一致（图3B、C）。例如，与其他方法的结果相比，海马角的锥体层（CAsp，簇10）和等皮质层（簇1、2、5和19）在MM检验结果中更清晰，空间连续性也更强。利用全部20张切片的数据，所有方法均能将大脑核团（CNU，簇11）与邻近区域（如纤维束（FT，簇3）、脑室

系统（VS，簇18）以及皮层下板（CTXsp，簇4）区分开来。丘脑（TH，簇8和13）和下丘脑（HY，簇6和14）也被清晰划分。值得注意的是，丘脑的两个亚区得到了良好分离：感觉运动皮层相关丘脑（DORsm，簇8）和多模态联合皮层相关丘脑（DORpm，簇13）。标记基因富集分析进一步表明，MM检验所得簇富集了区域特异性标记基因，显示出与已知神经解剖学注释的高度一致性（图3D、S5）。例如，同型皮层、下丘脑、丘脑和CNU簇均高表达各自的标记基因。

有趣的是，只有 MM检验能够清晰地勾勒出齿状回（DG）簇（簇 17），并将其与 CAsp 簇区分开来（图 3C）。标记基因分析证实，MM检验中的 DG 簇高表达 DG 特异性标记基因，而 CAsp 簇则表达 CA特异性标记（图3D、S6）。相比之下，其他方法未能生成 DG 或 CA 标记基因表达强烈且具有特异性的簇。这种区别在三维可视化中尤为明显（图4A）：MM检验显示 CAsp 簇与 DG 簇之间的边界清晰得多，与艾伦脑图谱中的三维注释高度吻合（图 4B、C）。其他方法生成的三维可视化进一步表明，它们无法区分 DG（图 4C）。

值得注意的是，簇 16 也表现出 DG 标记基因的高表达，可能代表一个额外的 DG 簇。二维可视化（图 3C）仅显示两个属于簇 16 的点，而三维可视化显示簇 16 和 17 相邻，二者可能共同代表 DG（图 4D）。此外，与其他方法生成的可视化相比，MM 检验识别出的同型皮层簇的层状结构在三维可视化中显得清晰得多。

与其他方法生成的结果相比。Moran、Binspect 和 SCFS 等方法错误地将同型皮层簇中的许多点归类，并错误地纳入了许多可能属于 CTXsp 的点，这表明它们的空间特异性较低（图 4E）。

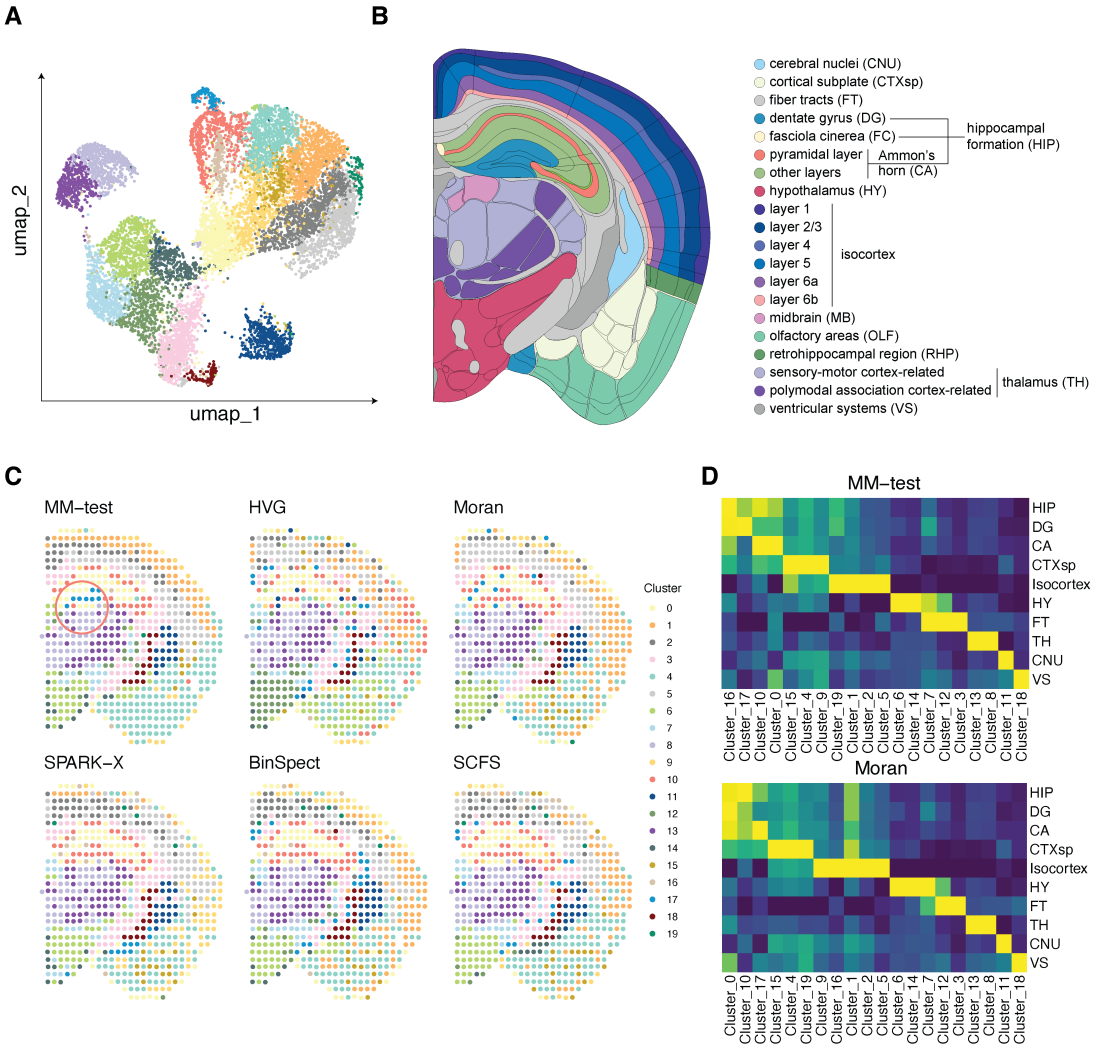

图3：（A）采用所提出的MM检验方法识别的细胞簇的UMAP可视化。（B）基于艾伦脑图谱的脑区注释。（C）使用各种方法在全部20张切片中筛选出的空间变异基因，对切片23B进行聚类的结果。红圈标出齿状回区域，该区域只有MM检验能够清晰识别。箭头指向齿状回区域。（D）标记基因富集分析：热图展示了MM检验和Moran识别的簇中区域特异性标记基因的表达模式。标记基因取自艾伦脑图谱（Lein等人，2007）。脑区缩写定义见（B）。

随后，我们通过改变纳入分析的切片数量进行了敏感性分析。具体而言，我们将切片数量从 20 减少到 12、5 和 1，并重复相同的分析流程。MM检验始终能识别出精细的簇，例如

CAsp、齿状回和 CNU 簇（表3 和图 S7–S9），在使用 20 至 5 张切片的数据时均能识别出来。相比之下，其他方法在切片数少于 20 时无法检测到这些簇。仅使用单张切片的数据时，MM检验和其他竞争方法都无法清晰地区分齿状回和 CNU 簇与其相邻区域，这凸显了纳入三维多切片空间转录组数据的优势。

最后，为了评估不同方法进行空间变异基因筛选的准确性，我们将每种方法排名靠前的 K 选定基因与从艾伦脑图谱（Lein等人，2007）整理出的12个脑区中的已知标记基因进行了比较。与其他竞争方法相比，MM检验在不同的 K选择下始终筛选出更多已知标记基因。例如，在CAsp和齿状回区域，MM检验鉴定出的标记基因数量约为其他方法的1.4倍（图5），从而改善了这些解剖学上具有区别的区域中的聚类表现。在包含12个、5个和1个切片的数据集中也观察到了这一优势，但程度有所降低（图 S10–S12）。这些结果表明，MM检验卓越的聚类表现可能归因于其更高的空间变异基因筛选准确性。

8 讨论

空间变异基因筛选是空间转录组数据分析中的一个基础问题。为解决这一问题，我们开发了MM检验，这是一种利用辅助信息筛选空间变异基因的无分布假设方法。在空间组学应用中，我们将辅助信息表示为距离矩阵，用于刻画组织位置之间的物理邻近性，使MM检验能够无缝应用于二维和三维空间转录组数据集。

不同于主要依赖基因表达的空间自相关性的现有方法，MM检验通过比较未知空间簇之间的表达均值来识别空间变异基因

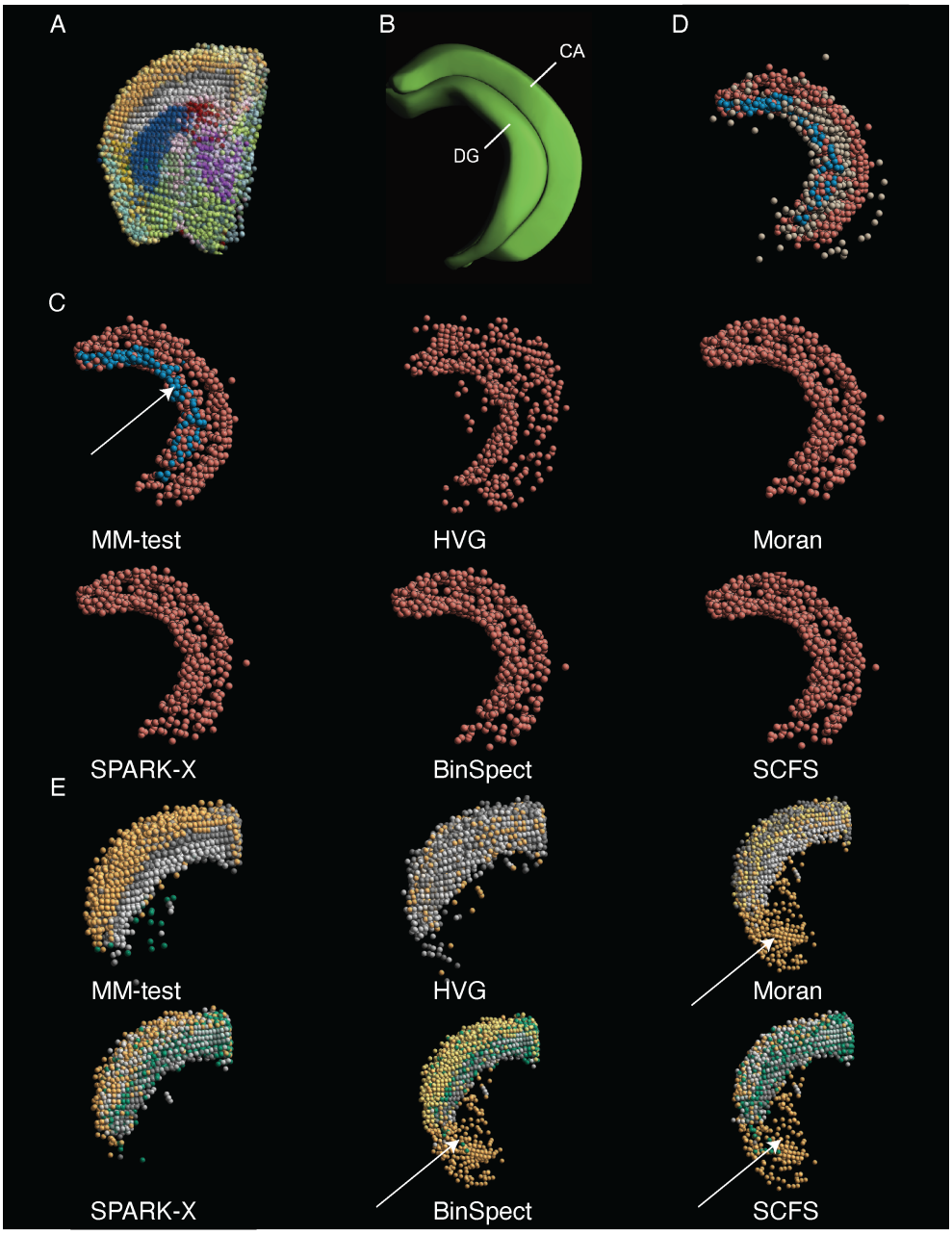

图4：MM检验在20个连续切片中识别出的簇的三维可视化。(A) MM检验识别出的所有簇。(B) 海马角和齿状回的三维解剖学注释。(C) 所有方法识别出的CAsp和齿状回区域（簇10和17）。(D) 所有方法中由MM检验识别出的CAsp（簇10）和齿状回区域（簇16和17）。(E) 所有方法识别出的新皮质区域。箭头指向来自CTXsp区域的误分类空间点（即Moran簇1）。

，从而能够稳健地检测与复杂空间域结构相关的空间变异基因。从理论上看，我们证明了MM检验能够以较高概率区分簇间具有异质均值的特征与具有同质均值的特征，证实了其筛选空间域间表达不同的基因的能力。进一

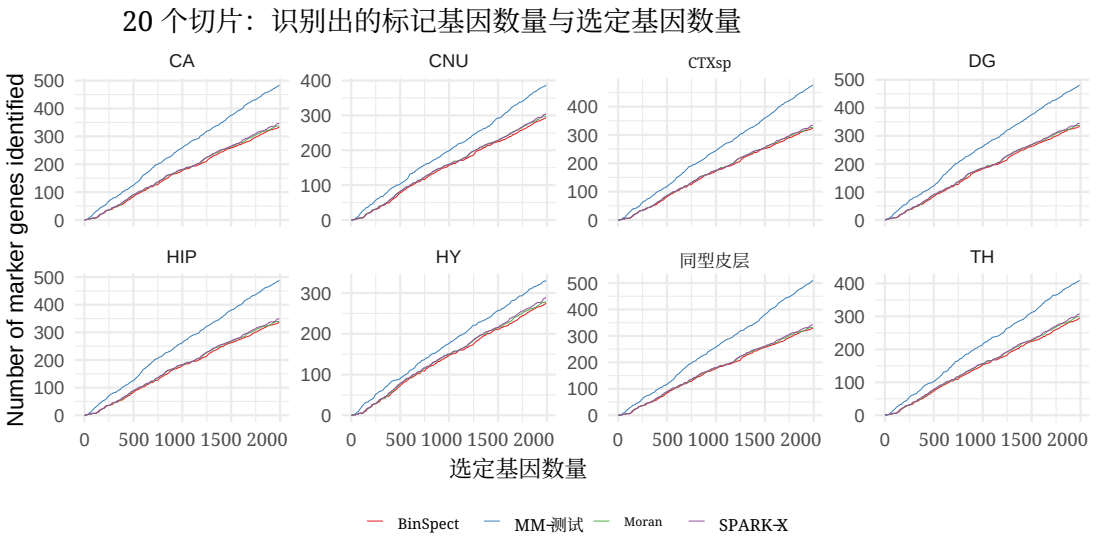

图 5：不同方法识别出的排名前列的空间变异基因中，各区域标记基因的数量。缩写定义见图3（B）。

表 3：切片数量变化时的敏感性分析。对勾（√）表示相应空间域已由给定方法成功识别，叉号（）表示未检测到该空间域。

切片数量	空间域	MM检验	HVG	Binspect	Moran	SPARK-X	SCFS
20	CAsp和齿状回	✓	✗	✗	✗	✗	✗
	CNU	✓	✗	✓	✓	✓	✓
	DORsm和DORpm	✓	✓	✓	✓	✓	✓
12	CAsp和齿状回	✓	✗	✗	✗	✗	✗
	CNU	✓	✗	✗	✗	✗	✗
	DORsm和DORpm	✓	✓	✓	✓	✓	✓
5	CAsp 和齿状回	✓	✗	✗	✗	✗	✗
	CNU	✓	✗	✗	✗	✗	✗
	DORsm 和 DORpm	✗	✗	✗	✗	✗	✗

此外，我们为后选择聚类建立了误差界，证明基于 MM检验选定特征的聚类，其准确度与使用真实聚类相关特征所获得的准确度相当。因此，在实际的空间转录组应用中，MM检验能够更准确地检测与空间域相关的标记基因，并促进更精确的空间域识别。

空间转录组数据集通常附有组织学图像，可从中提取额外的辅助信息。利用这些信息，MM检验可以扩展到三维多切片数据集之外，直接从包含多个互不相关切片的数据集中筛选空间变异基因。在其他应用中，例如多组学整合，距离矩阵

可以根据互补数据模态（如蛋白质组学、代谢组学或表观基因组学）构建，使MM检验能够用于空间转录组学以外的特征选择。更广泛地说，对于能够将样本间关系信息编码为距离的应用，MM检验提供了一个通用的特征筛选框架。

当主要目标是识别空间域的标记基因时，将 MM 检验与 knockoff 方法结合，可为传统的两步方法提供一种具有严格统计学依据的替代方案。传统方法先进行空间聚类，再进行差异表达分析，存在“重复使用数据”的问题：使用同一数据进行聚类和检验会导致 p 值无效，并使错误发现率升高（Chen 等人，2024年）。相比之下，MM 检验无需依赖预先聚类，即可直接筛选空间信息基因，并严格控制错误发现率。虽然最初无法确定这些已识别标记所对应的具体空间域，但可以事后将其归入表达水平最高的簇。此分配步骤不涉及进一步检验，因此能够保持标记选择的统计有效性。

9 数据可用性

R软件包 `MMtestSVG`及其配套代码和数据集可在 Zenodo 上获取，DOI: 10.5281/zenodo.17070739。

参考文献

Banerjee, T., G. Mukherjee, and P. Radchenko（2017年）。大规模聚类分析中的特征筛选。*Journal of Multivariate Analysis 161*，191–212。

Butler, A., P. Hoffman, P. Smibert, E. Papalexi 和 R. Satija（2018 年）。整合单

细胞转录组数据，以适用于不同条件、技术和物种。《自然-生物技术》36(5), 411–420. Candes, E., Y. Fan, L. Janson 和 J. Lv（2018 年）。淘金：用于高维受控变量选择的“model-x”knockoff。《皇家统计学会杂志 B辑：统计方法学》80(3), 551–577.

Chan, Y.-b. 和 P. Hall（2010年）。利用混合群体的证据为超高维数据聚类选择变量。《美国统计学会杂志》*105*(490), 798–809。

Chen, A., S. Liao, M. Cheng, K. Ma, L. Wu, Y. Lai, X. Qiu, J. Yang, J. Xu, S. Hao 等（2022年）。利用 DNA纳米球图案化阵列绘制小鼠器官发生的时空转录组图谱。《细胞》*185*(10), 1777–1792。

Chen, C., H. J. Kim 和 P. Yang（2024年）。评估用于空间转录组数据的空间变异基因检测方法。*Genome Biology 25*(1), 18。

Chen, X., Q. Ran, J. Tang, Z. Chen, S. Huang, X. Shi 和 R. Xi（2025 年）。空间转录组学中空间变异基因识别算法的基准评估。***Bioinformatics***, btaf131。

Chen, Z.、C. Wang、S. Huang、Y. Shi和R. Xi（2024年）。直接选择细胞类型标记基因用于单细胞聚类分析。《细胞报告方法》*4*(7)。Dries, R.、Q. Zhu、R. Dong、C.-H. L. Eng、H. Li、K. Liu、Y. Fu、T. Zhao、A. Sarkar、F. Bao等（2021 年）。Giotto：用于空间表达数据整合分析与可视化的工具箱。*Genome Biology 22*, 1–31。Fan, J.和J. Lv（2008年）。超高维特征空间的确定性独立筛选。《皇家统计学会会刊：*B*辑（统计方法）》*70*(5), 849–911。

Hao, M.、K. Hua和X. Zhang（2021 年）。SOMDE：一种利用自组织映射识别空间变异基因的可扩展方法。《生物信息学》*37*(23)，4392–4398。Hu, J.、X. Li、K. Coleman、A. Schroeder、N. Ma、D. J. Irwin、E. B. Lee、R. T. Shinohara和M. Li（2021 年）。SpaGCN：整合基因表达、空间位置和组织学，通过图卷积网络识别空间域和空间变异基因。*Nature Methods 18*(11)，1342–1351。Hunter, D. R.和K. Lange（2004年）。MM算法教程。《美国统计学家》*58*(1)，30–37。Jin, J.和W. Wang（2016年）。面向高维聚类的影响特征主成分分析。《统计年鉴》*44*(6)，2323–2359。Lein, E. S.、M. J. Hawrylycz、N. Ao、M. Ayres、A. Bensinger、A. Bernard、A. F. Boe、M. S. Boguski、K. S. Brockway、E. J. Byrnes等（2007年）。成年小鼠脑基因表达的全基因组图谱。《自然》*445*(7124)，168–176。Li, Q.、M. Zhang、Y. Xie和G. Xiao（2021 年）。基于高斯过程的空间分子图谱数据贝叶斯建模。《生物信息学》*37*(22)，4129–4136。Liu, J.、W. Zhong和R. Li（2015年）。超高维数据特征筛选的选择性综述。《中国科学：数学》*58*(10)，1–22。Liu, T.、Y. Lu、B. Zhu和H. Zhao（2022年）。通过特征选择对高维数据进行聚类。《生物统计学》。Marx, V.（2021 年）。年度方法：空间分辨转录组学。*Nature Methods 18*(1)，9–14。Maynard, K. R.、L. Collado-Torres、L. M. Weber、C. Uytingco、B. K. Barry、S. R. Williams、J. L. Catallini、M. N. Tran、Z. Besich、M. Tippani等（2021 年）。人类背外侧前额叶皮层中转录组尺度的空间基因表达。*Nature Neuro-*

《科学》*24*(3), 425–436. Miller, B. F., D. Bambah-Mukku, C. Dulac, X. Zhuang, and J. Fan (2021 年). 利用非均匀细胞密度对空间分辨单细胞转录组数据中的空间基因表达异质性进行表征。《基因组研究》*31*(10), 1843–1855. Ortiz, C., J. F. Navarro, A. Jurek, A. M¨artin, J. Lundeberg, and K. Meletis (2020 年). 成年小鼠脑的分子图谱。《科学进展》*6*(26), eabb3446. Qian, X., K. Coleman, S. Jiang, A. J. Kriz, J. H. Marciano, C. Luo, C. Cai, M. D. Manam, E. Caglayan, A. Lai, et al. (2025 年). 空间转录组学揭示人类皮层层次与区域特化。《自然》, 1–11. Sarkar, A. K. and M. Stephens (2020 年). 分离测量模型与表达模型，厘清单细胞RNA测序分析中的混淆。《自然遗传学》*53*, 770 – 777. Schott, M., D. Le´on-Peri˜n´an, E. Splendiani, L. Strenger, J. R. Licha, T. M. Pentimalli, S. Schallenberg, J. Alles, S. S. Tagliaferro, A. Boltengagen, et al. (2024年). Open-ST：三维高分辨率空间转录组学。《细胞》*187*(15), 3953–3972. Sharma, A., E. Kosasih, J. Zhang, A. Brintrup, and A. Calinescu (2022年). 数字孪生：最新进展、理论与实践、挑战及开放研究问题。《工业信息集成杂志》*30*, 100383. Shi, H., Y. He, Y. Zhou, J. Huang, K. Maher, B. Wang, Z. Tang, S. Luo, P. Tan, M. Wu, et al. (2023年). 分子分辨率下的小鼠中枢神经系统空间图谱。《自然》*622*(7983), 552–561. Srivatsan, S. R., M. C. Regier, E. Barkan, J. M. Franks, J. S. Packer, P. Grosjean, M. Duran, S. Saxton, J. J. Ladd, M. Spielmann, et al. (2021 年). 胚胎尺度的单细胞空间转录组学。《科学》*373*(6550), 111–117. Su, H. and Y. Cui (2025 年). 旋转不变性对于准确检测空间转录组学中的空间变异

基因至关重要。*Nature Communications 16*(1), 7122. Sun, S., J. Zhu, and X. Zhou (2020 年). 空间分辨转录组学研究中空间表达模式的统计分析。*Nature Methods 17*(2), 193–200. Svensson, V., S. A. Teichmann, and O. Stegle (2018 年). SpatialDE：空间变异基因的识别。*Nature Methods 15*, 343 – 346.

Tang, J., Z. Chen, K.Qian, S. Huang, Y. He, S.Yin, X. He, B. Ye, Y. Zhuang, H.Meng, 等 (2026)。可解释的多模态降维框架 spahdmap 提升了空间转录组学的分辨率。*Nature Cell Biology*，1–15。

Wang, C., Z. Chen, 和 R. Xi (2025)。用于计数数据聚类分析的特征筛选及其在单细胞 RNA 测序中的应用。*The Annals of Applied Statistics 19*(4)，2738–2758。

Wedderburn, R. W. (1974)。准似然函数、广义线性模型与高斯—牛顿法。*Biometrika 61*(3)，439–447。

Yuan, X., Y. Ma, R. Gao, S. Cui, Y. Wang, B. Fa, S. Ma, T. Wei, S. Ma, 和 Z. Yu (2024 年)。HEARTSVG：一种快速、准确的识别大规模空间转录组学中空间变异基因的方法。*Nature Communications 15*(1)，5700。

Zhang, M., X. Pan, W. Jung, A. R. Halpern, S. W. Eichhorn, Z. Lei, L. Cohen, K. A. Smith, B. Tasic, Z. Yao, H. Zeng 和 X. Zhuang（2023）。全小鼠脑分子定义且空间分辨的细胞图谱。《自然》*624*，343–354。

Zhu, J., S. Sun 和 X. Zhou（2021 年）。SPARK-X：非参数建模使大规模空间转录组研究中的空间表达模式检测具备可扩展性与稳健性。*Genome Biology 22*(1)，184。