文章编号:1007－6735(2024)06－0698－10　DOI: 10.13255/j.cnki.jusst.20230812003

空间解析单细胞转录组的优化算法研究

皇站飞，　赵桂华
（上海理工大学 理学院，上海 200093）

摘要：由于单细胞和空间转录组技术都存在一定的不足，整合单细胞转录组和空间转录组技术应运而生。为提高单细胞矩阵到空间矩阵映射的相似度，降低损失函数值，通过改进深度学习 Tangram 算法的目标函数，同时受龙格库塔方法的启发对优化算法 Adam 的梯度值进行修正，开发了 RK-Tangram 算法。将其应用到 3 组模拟数据与真实的小鼠大脑皮质、运动和视觉区域的数据上，与原始 Tangram 算法相比，结果表明，RK-Tangram 算法不仅提高了映射的相似度，降低了损失函数值，而且扩展了空间转录组的全基因组图谱，并纠正了低质量的空间测量。另外，通过解卷积将空间转录组数据转化为单细胞数据，提供了一个更高分辨率的组织类型图谱。

关键词：深度学习；梯度下降；解卷积；转录组
中图分类号：Q 503　　**文献标志码：**A

Research on the optimization algorithm for spatially resolved single-cell transcriptomes

HUANG Zhanfei，　ZHAO Guihua
(*College of Science, University of Shanghai for Science and Technology, Shanghai 200093, China*)

Abstract: As both single-cell and spatial transcriptome techniques have certain shortcomings, techniques integrating single-cell transcriptome and spatial transcriptome were developed. In order to improve the similarity of mapping single-cell matrix to spatial matrix and reduce the loss function value, the RK-Tangram algorithm was developed by improving the objective function of the Tangram algorithm for deep learning and also correcting the gradient value of the optimization algorithm Adam, inspired by the Runge-Kutta method. Applying it to three sets of simulated data and real data from mouse brain cortical, motor and visual regions, compared with the original Tangram algorithm, the

收稿日期: 2023−08−12
基金项目: 国家自然科学基金资助项目 (61807017)
第一作者: 皇站飞(1998−)，男，硕士研究生．研究方向：生物信息学．E-mail：huang_zfy@163.com
通信作者: 赵桂华(1981−)，女，副教授．研究方向：生物信息学、随机控制．E-mail：zgh-hit1108@163.com

引文格式: 皇站飞，赵桂华．空间解析单细胞转录组的优化算法研究[J]．上海理工大学学报，2024，46(6)：698-707.
Citation: HUANG Zhanfei, ZHAO Guihua. Research on the optimization algorithm for spatially resolved single-cell transcriptomes[J]. Journal of University of Shanghai for Science and Technology, 2024, 46(6)：698-707.

results showed that the RK-Tangram algorithm not only improved the similarity of the mapping and reduced the loss function value, but also extended the genome-wide mapping of the spatial transcriptome and corrected spatial measurements with low-quality. In addition, deconvoluting the spatial transcriptome's data to single cell's provided a higher resolution mapping of tissue types.

Keywords: *deep learning*; *gradient descent*; *deconvolution*; *transcriptome*

近年来，单细胞和空间转录组分析得到迅速发展——单细胞转录组测序（如 scRNA-seq[1]）、空间转录组技术（如 ST/Visium[2]，Slide-seq[3]，Slide-seqV2[4] 和 HDST[5]）以及靶向原位捕获技术（如 MERFISH[6-7]，smFISH[8]，osmFISH[9]，STARmap[10]，SeqFISH[11-12]，seqFISH+[13]），这些技术的进步为高分辨率空间图谱的绘制开辟了道路[14]。单细胞转录组测序通过原位杂交和测序，在单细胞分辨率的水平上解析转录组，却丢失了空间位置信息[2]。空间转录组技术是基于空间条码和测序的原位捕获技术，在空间上解析转录组，对整个转录组进行空间条形码标记，但捕获率有限，且空间分辨率大于单细胞水平（提高分辨率又会造成大量信息丢失）[15]。靶向原位测序可在单细胞分辨率下生成固定细胞或组织的多重表达谱，它首先将 mRNA 原位逆转录成 cDNA，再通过锁式探针（padlock probe）开展靶点识别和滚环扩增（RCA）。虽然靶向原位捕获技术解决了空间位置信息丢失和分辨率低的问题，但受到通量的限制，每次测量的基因数仅数百个，如果增加探针的数量又会影响基因的准确性[10]。

目前也有了一些整合单细胞数据与空间转录组数据的方法：Cell2location 方法[16] 通过集成单细胞和空间转录组数据，以全面绘制组织细胞结构的贝叶斯模型；SPOTlight 方法[17] 使用非负矩阵分解和解卷积的方法，将 ST 与 scRNA-seq 数据结合，推断出复杂组织中细胞类型与位置信息。Tangram 算法[18] 通过单细胞表达矩阵预测空间表达矩阵，使用余弦相似度衡量预测的准确性。

Tangram 算法是一种深度学习算法，在单细胞分辨率上学习转录组的空间基因表达图，并将这些图与来自同一标本的组织学和解剖学信息联系起来。通过单细胞矩阵与深度学习得到的映射矩阵来预测空间矩阵，并使用余弦相似度衡量预测空间矩阵与期望空间矩阵的相似度，使用损失函数衡量算法的整体损失大小，这里映射矩阵是通过 Adam[19] 优化器进行深度学习获得的。为了提高空间预测的准确性，更好地预测空间细胞类型，对损失函数进行改进。同时，受龙格库塔方法的启发，对 Adam 中梯度值的计算进行线性加权修正，能够提高当前时刻梯度值的可信度，因此，本文开发了 RK-Tangram 算法。在给定的 3 组模拟数据集上，与 Tangram 算法相比，RK-Tangram 算法表现出更好的鲁棒性，如：对不同类型的数据集，RK-Tangram 算法输出更低的损失值和更高的余弦相似度。最后使用 RK-Tangram 分析 3 组真实数据：小鼠大脑皮质数据、运动和视觉区域的单细胞转录组数据，及来自上述 3 组数据的同一组织切片的空间转录组 Visium，Slide-seq 和 MERFISH 数据集。与 Tangram 算法相比，RK-Tangram 收敛速度更快，预测更精准，预测的空间矩阵与期望的空间矩阵相似度更高，且解卷积的细胞类型分层更加明显，更有助于生物学与病理学的研究，促进新的发现。

1　RK-Tangram 算法

本文的主要目标是寻找映射矩阵 $\hat{\boldsymbol{M}}$，对 $\hat{\boldsymbol{M}}$ 进行深度学习，使目标损失函数值达到最小。为了提高预测空间矩阵的准确性，确保每个细胞的空间概率尽可能集中在一个点，以及使细胞过滤更加精准，降低混乱程度，构造损失函数如下：

$$F\left(\hat{\boldsymbol{M}},\hat{\boldsymbol{f}}\right)=\zeta_{\mathrm{kl}}\left(\boldsymbol{m}^{f},\boldsymbol{d}\right)-\frac{1}{n_{\mathrm{genes}}}\sum_{k}^{n_{\mathrm{genes}}}\cos\langle(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{S}^{f})_{*,k},\boldsymbol{G}_{*,k}\rangle-$$
$$\frac{1}{n_{\mathrm{voxels}}}\sum_{j}^{n_{\mathrm{voxels}}}\cos\langle(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{S}^{f})_{j,*},\boldsymbol{G}_{j,*}\rangle+$$
$$\left(\frac{1}{n_{\mathrm{cells}}\times n_{\mathrm{voxels}}}\sum_{i,j}^{n_{\mathrm{cells}},n_{\mathrm{voxels}}}M_{ij}\log M_{ij}\right)^{2}+$$
$$\lambda_{1}\left|\sum_{i}^{n_{\mathrm{cells}}}f_{i}-n_{\mathrm{target_{cells}}}\right|-\lambda_{2}\sum_{i}^{n_{\mathrm{cells}}}(f_{i}\log f_{i})$$

式中：$\boldsymbol{M}$ 表示单细胞到空间变换的 $n_{\mathrm{cells}}\times n_{\mathrm{voxels}}$ 映

射矩阵，元素M_{ij}表示细胞i属于空间网格点j的概率，有$\sum_{j}^{n_{\text{voxels}}} M_{ij}=1$，其中$\boldsymbol{M}$是通过对$\hat{\boldsymbol{M}}$做softmax变换[20]获得，即$M_{ij}=\mathrm{e}^{\hat{M}_{ij}}\Big/\left(\sum_{k=1}^{n_{\text{voxels}}}\mathrm{e}^{\hat{M}_{ik}}\right)$，$\boldsymbol{M}^{\mathrm{T}}$为$\boldsymbol{M}$的转置；$\hat{\boldsymbol{f}}=\left(\hat{f}_1,\hat{f}_2,\cdots,\hat{f}_{\text{cells}}\right)$表示$n_{\text{cells}}$维的过滤向量，$\boldsymbol{f}=(f_1,f_2,\cdots,f_{\text{cells}})$为对$\hat{\boldsymbol{f}}$执行sigmoid变换[21]获得的结果，使得$0\leqslant f_i\leqslant 1$，$\hat{f}_i$表示细胞$i$是否被过滤掉，为布尔值，$\hat{f}_i=1$表示细胞$i$被保留用于映射，$\hat{f}_i=0$表示细胞$i$被过滤掉。$\boldsymbol{m}^{\boldsymbol{f}}=\left(m_1^f,m_2^f,\cdots,m_{n_{\text{voxels}}}^f\right)$表示细胞过滤后预测空间网格点的细胞密度；$\boldsymbol{d}=\left(d_1,d_2,\cdots,d_{n_{\text{voxels}}}\right)$表示先验细胞密度，满足$\sum_{i}^{n_{\text{voxels}}} d_i=1$，且$0\leqslant d_i\leqslant 1$；$\boldsymbol{S}$表示$n_{\text{cells}}\times n_{\text{genes}}$单细胞表达矩阵，其中元素$S_{jk}\geqslant 0$为单细胞$j$中基因$k$的表达水平，$n_{\text{genes}}$为基因数，$n_{\text{cells}}$为单细胞数；$\boldsymbol{G}$表示$n_{\text{voxels}}\times n_{\text{genes}}$空间表达矩阵，其中元素$G_{ik}\geqslant 0$为空间网格$i$中基因$k$的表达水平，$n_{\text{voxels}}$为网格点数；$\boldsymbol{M}^{\mathrm{T}}\boldsymbol{S}^{\boldsymbol{f}}$表示预测空间矩阵，$\lambda_1$，$\lambda_2$为权重系数，默认为1；$m_j^f=\sum_{i}^{n_{\text{cells}}} M_{ij}^f\Big/\sum_{i}^{n_{\text{cells}}} f_i$表示网格点$j$的预测细胞密度；$\boldsymbol{S}^{\boldsymbol{f}}=\mathrm{diag}(\boldsymbol{f})\cdot\boldsymbol{S}$为过滤后的单细胞数据，$\boldsymbol{M}^{\boldsymbol{f}}=\mathrm{diag}(\boldsymbol{f})\cdot\boldsymbol{M}$为过滤后的映射矩阵，$n_{\text{target}_{\text{cells}}}$为希望过滤后的细胞数量，若不给定即默认为细胞分割数量，记Exp为点中基因或细胞类型表达量。损失函数第一项ζ_{kl}为kl散度，衡量期望细胞密度与预测细胞密度的相似度，kl散度值越小说明两分布之间越相似；第二项$\cos\langle(\boldsymbol{M}^T\boldsymbol{S}^f)_{*,k},\boldsymbol{G}_{*,k}\rangle$为映射得到的空间矩阵第$k$列与真实空间表达矩阵第$k$列夹角的余弦，即为体素网格中基因预测值与期望值的余弦相似度；第三项同理；第四项为映射矩阵的信息熵，确保每个细胞的空间概率尽可能集中在一个点；第五项为常数项，确保过滤后细胞数量与预先给定的目标细胞数量一致；最后一项为细胞过滤信息熵。

RK-Tangram对Tangram损失函数进行了两方面的改进：对Tangram损失函数中熵项取均值并平方；将Tangram算法的损失函数中常数项改进为过滤向量信息熵。在处理高通量数据集时，改进后RK-Tangram算法的损失函数中余弦相似度灵敏度较高，损失函数较小时，余弦相似度收敛于较大值，预测的准确性高，避免了Tangram算法因高通量数据引起余弦相似度灵敏度低的不足；常数项改进为过滤向量信息熵，这提高了细胞过滤的准确性。

本文的RK-Tangram算法不仅对Tangram损失函数进行了改进，而且对求映射矩阵$\hat{\boldsymbol{M}}$的迭代算法进行了改进。龙格库塔算法是用于求非线性常微分方程数值近似解的重要方法，主要是对斜率的矫正，使得预测值更接近于真实值，提高预测精度。根据该思想，对迭代算法Adam中的梯度值进行改进，用于更新映射矩阵$\hat{\boldsymbol{M}}$。

简要描述RK-Tangram优化算法(当损失函数不包含过滤向量$\hat{\boldsymbol{f}}$，即只包含变量$\hat{\boldsymbol{M}}$)的更新流程：

输入 β_1，β_2，N，ε，α；输出 $\hat{\boldsymbol{M}}_{\text{t}}$

初始化 $\begin{cases}\hat{\boldsymbol{m}}_0=0,\hat{\boldsymbol{v}}_0=0,t=1,\\ \hat{\boldsymbol{M}}_0=\text{random.normal()}\end{cases}$

(a) $\tilde{\boldsymbol{g}}_t\leftarrow\nabla_M F\left(\hat{\boldsymbol{M}}_{t-1}\right)$；

(b) $\hat{\boldsymbol{m}}_t\leftarrow\beta_1\hat{\boldsymbol{m}}_{t-1}+(1-\beta_1)\tilde{\boldsymbol{g}}_t$；

(c) $\tilde{\boldsymbol{v}}_t\leftarrow\beta_2\hat{\boldsymbol{v}}_{t-1}+(1-\beta_2)\tilde{\boldsymbol{g}}_t^2$；

(d) $p\leftarrow\alpha/(1-\beta_1^t)$；

(e) $\tilde{\boldsymbol{V}}_t\leftarrow\dfrac{\sqrt{\tilde{\boldsymbol{v}}_t}}{\sqrt{1-\beta_2^t}}+\varepsilon\boldsymbol{I}$；

(f) $\tilde{\boldsymbol{M}}_t\leftarrow\hat{\boldsymbol{M}}_{t-1}-p\dfrac{\tilde{\boldsymbol{m}}_t}{\tilde{\boldsymbol{V}}_t}$；

(g) $\boldsymbol{g}_t\leftarrow\dfrac{\nabla_M F\left(\tilde{\boldsymbol{M}}_t\right)+4\nabla_M F\left(\left(\tilde{\boldsymbol{M}}_t+\hat{\boldsymbol{M}}_{t-1}\right)/2\right)+\nabla_M F\left(\hat{\boldsymbol{M}}_{t-1}\right)}{6}$；

(h) $\hat{\boldsymbol{m}}_t\leftarrow\beta_1\hat{\boldsymbol{m}}_{t-1}+(1-\beta_1)\boldsymbol{g}_t$；

(i) $\hat{\boldsymbol{v}}_t\leftarrow\beta_2\hat{\boldsymbol{v}}_{t-1}+(1-\beta_2)\boldsymbol{g}_t^2$；

(j) $\boldsymbol{V}_t\leftarrow\dfrac{\sqrt{\hat{\boldsymbol{v}}_t}}{\sqrt{1-\beta_2^t}}+\varepsilon\boldsymbol{I}$；

(k) $\hat{\boldsymbol{M}}_t=\hat{\boldsymbol{M}}_{t-1}-p\dfrac{\hat{\boldsymbol{m}}_t}{\boldsymbol{V}_t}$；

(l) $t=t+1$；

(m) $t=N$时，输出$\hat{\boldsymbol{M}}_t$，否则转到(a)。

这里$F\left(\hat{\boldsymbol{M}}_t\right)$只包含损失函数前四项，$\nabla_M F$表示$F$关于$\boldsymbol{M}$的梯度，即$\tilde{\boldsymbol{g}}_t$；$\tilde{\boldsymbol{m}}_t$和$\tilde{\boldsymbol{v}}_t$为一阶动量和二阶动量；$\tilde{\boldsymbol{V}}_t$，$\tilde{\boldsymbol{M}}_t$和$\boldsymbol{V}_t$为中间过渡变量；$\boldsymbol{I}$为元素全为1的$n_{\text{cells}}\times n_{\text{voxels}}$矩阵；$\alpha$为步长，$p$为修正后的步长；$\boldsymbol{g}_t$为当前迭代下最终的梯度值；$\hat{\boldsymbol{m}}_t$和$\hat{\boldsymbol{v}}_t$为修正后的一阶动量和二阶动量；$\hat{\boldsymbol{M}}_t$为更新的映射矩阵。在算法运行中，令权重系数$(\beta_1,\beta_2)=(0.6,0.7)$，非零常数$\varepsilon=10^{-8}$，$\alpha=0.2$；给定迭代次数，并要求每100次输出结果。

使用余弦相似度和损失函数值评估预测空间矩阵的准确性，并对预测空间矩阵可视化，直接展示RK-Tangram算法的生物学意义。

2 结果与讨论

2.1 模拟数据结果

对真实数据进行分布拟合，发现数据分布接

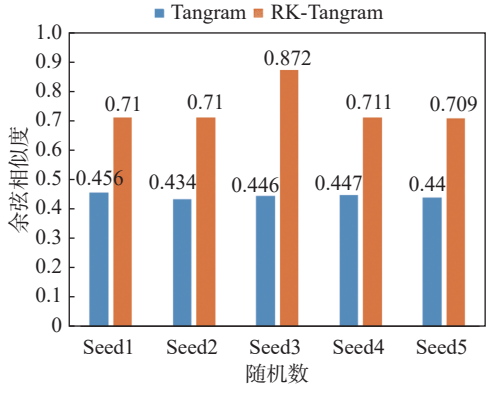
近于正态分布，于是本文通过软件包 numpy[22] 随机生成 3 组正态分布数据。第一组数据集，单细胞数据包含 300 个细胞和 5000 个基因的表达值；空间数据包含 200 个网格点和 5000 个基因的表达值。为了验证算法的鲁棒性，在保证正态分布期望 0.5 和方差 1 不变的情况下，使用不同的随机种子(seed)产生 5 组相同通量的数据集。第二组数据集，单细胞数据包含 400 个细胞和 10000 个基因的表达值；空间数据包含 300 个网格点和 10000 个基因的表达值，在保证正态分布期望值 0.2 和通量不变的情况下，改变方差大小，分别设置为 0.5，0.8，1.5 和 2。第三组数据集，单细胞数据包含 400 个细胞和 10000 个基因的表达值；空间数据包含 300 个网格点和 10000 个基因的表达值，在保证正态分布方差 0.5 和通量不变的情况下，改变期望大小，分别设置为 0 和 0.5。

余弦相似度越高说明预测的空间表达值与真实的空间表达值越吻合。对于第一组数据，从图 1(a)可知，RK-Tangram 算法的效果都一致优于 Tangram，RK-Tangram 算法的余弦相似度(score)一直稳定在 0.7 左右，Tangram 的余弦相似度收敛于 0.4 左右。对于第二和第三组数据集，由图 1(b)和图 1(c)可知，在方差和期望变化的情况下，RK-Tangram 算法的效果依旧高于 Tangram 算法。可见 RK-Tangram 算法对于不同类型的数据效果总是高于 Tangram 算法。从余弦相似度收敛趋势图 1(d)来看，RK-Tangram 算法在 50 步时基本收敛，而 Tangram 算法在 75 步时才收敛，这表明 RK-Tangram 收敛速度更快。

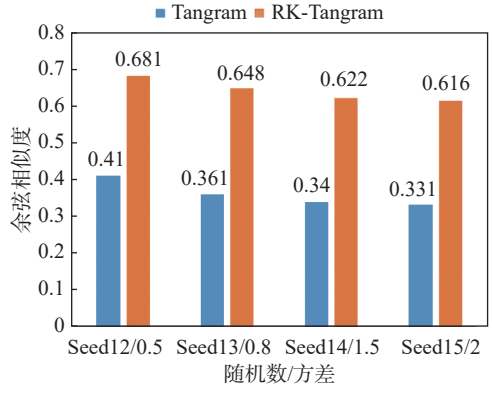
2.2　真实数据结果

2.2.1　成年小鼠大脑切片皮质层 scRNA-seq 单细胞数据与 Visium 空间表达数据

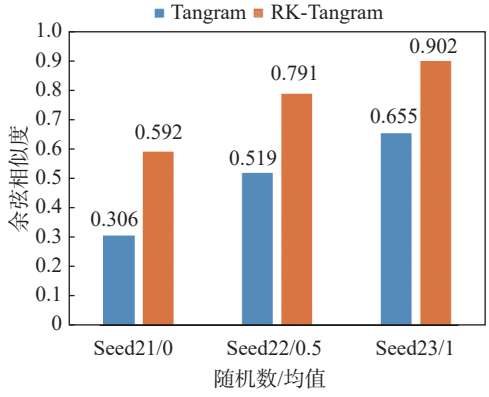
单细胞 scRNA-seq 和空间 Visium[2] 的成年小鼠大脑切片数据均通过 Squidpy[23] 下载。其中，单细胞 scRNA-seq 数据包含 21697 个单细胞和 36826 个基因；空间 Visium[2] 数据集提取于皮质层 1，2，3 和 4 的类数据，包含 324 个网格点和 16562 个基因。选择每个细胞类型特有的 1401 个基因作为标记基因，经过预处理，筛选出 1280 个标记基因作为训练基因。对单细胞数据的细胞类型绘制 UMAP 图像，从图 2(a)可知，数据具有很好的分层和聚类。



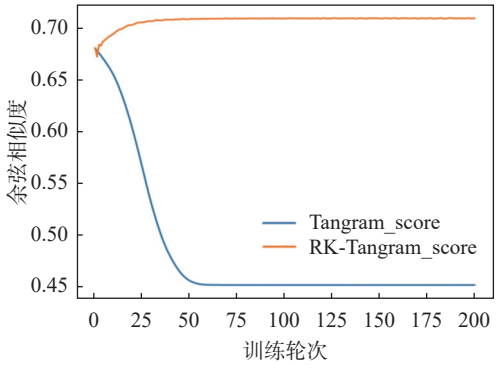
(a) 第一组数据集的余弦相似度收敛值对比图



(b) 第二组数据集的余弦相似度收敛值对比图



(c) 第三组数据集的余弦相似度收敛值对比图

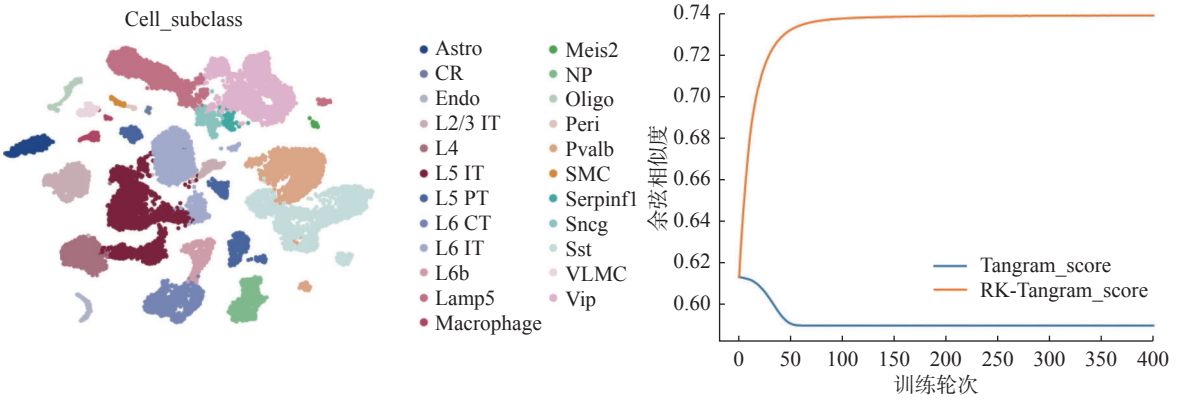


(d) 第一组数据集(seed1)的余弦相似度对比图

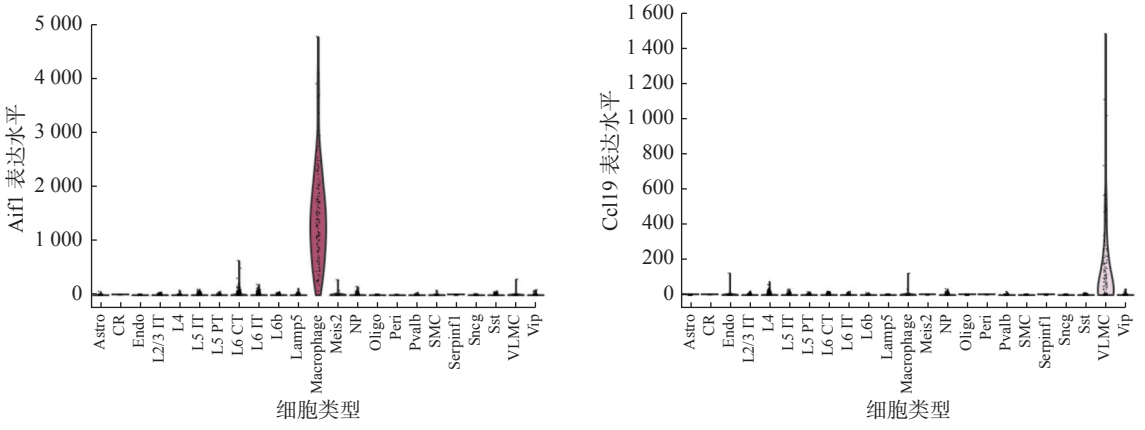
图 1　模拟数据的分析

Fig.1　Analysis of simulated data

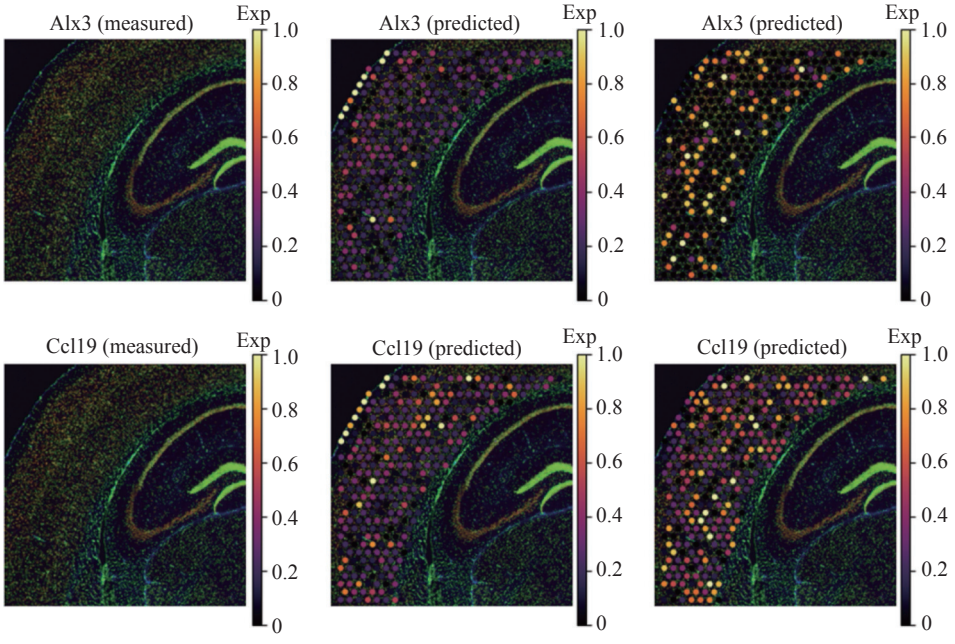
a. $\lambda_1 = \lambda_2 = 0$时，使用不带过滤项的损失函数计算映射矩阵，将单细胞数据的细胞类型映射到空间上，获得不同细胞类型的空间概率分布。与Tangram相比，RK-Tangram的余弦相似度更高(图2(b))；从生物学意义上来看，RK-Tangram扩展了全基因组，单细胞测序显示基因Alx3和Ccl19



(a) scRNA-seq 的 UMAP 图　　(b) 余弦相似度对比图



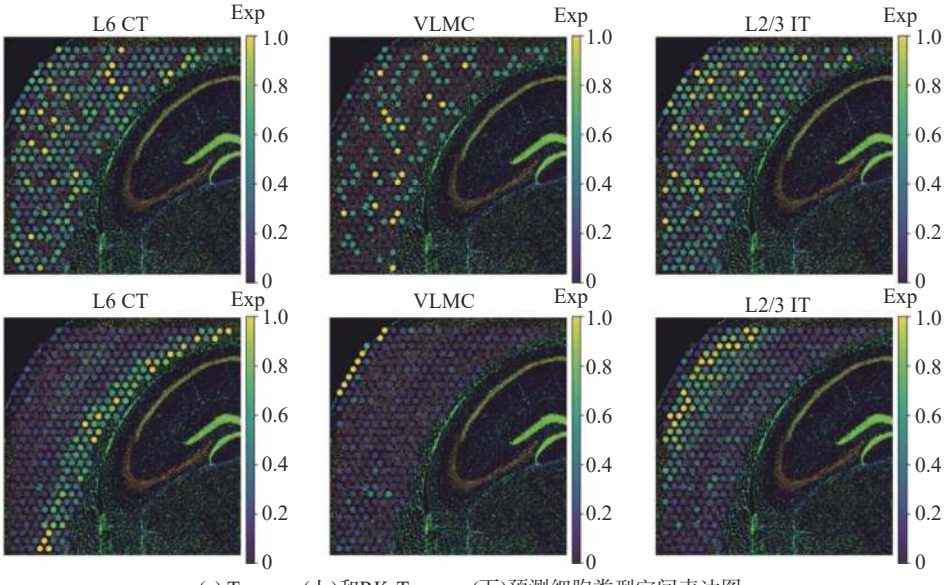
(c) 基因 Alx3 和 Ccl19 的小提琴图



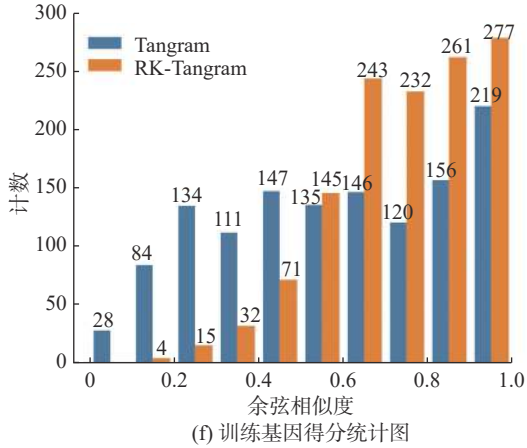
(d) 原始空间测序(左)、RK-Tangram(中)与Tangram(右)预测基因空间表达图

图 2　成年小鼠大脑皮质层数据分析

Fig.2　Analysis of adult mouse brain cortical layer data



(e) Tangram(上)和RK-Tangram(下)预测细胞类型空间表达图



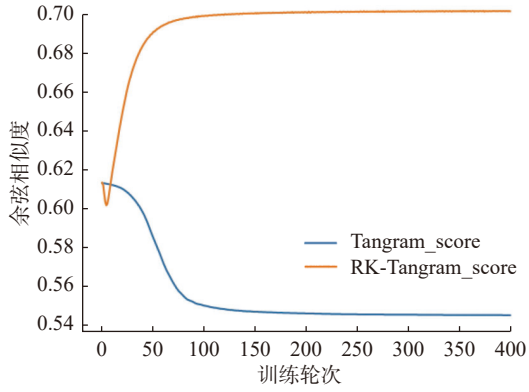
(f) 训练基因得分统计图

图 2　（续）

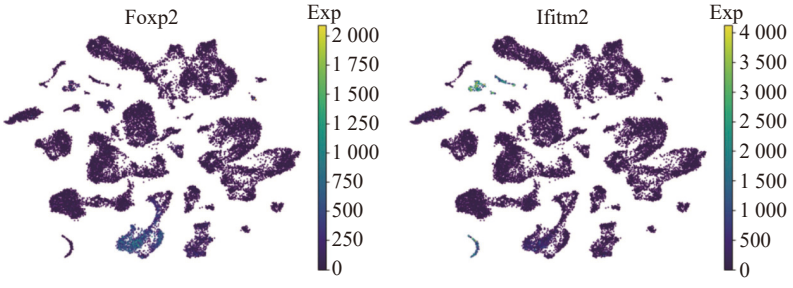
Fig.2　(continued)

在部分细胞类型中表达(图 2(c))，而在空间转录组测序中未检测到(图 2(d)左)。RK-Tangram 通过映射矩阵预测出了基因 Alx3 和 Ccl19 的表达(图 2(d)中)，且表达区域与单细胞测序一致，而 Tangram[18] 虽然预测出基因 Alx3 和 Ccl19 的表达，但表达区域与单细胞测序不一致(图 2(d)右)。为揭示细胞类型的空间分布特征，利用映射矩阵与单细胞的细胞类型注释矩阵来预测细胞类型的空间分布。从图 2(e)来看，RK-Tangram 预测的细胞类型空间表达分层更明显，说明预测的效果更好。另外由训练基因得分统计图 2(f)可知，RK-Tangram 的基因余弦相似度整体比 Tangram 算法的高，说明预测的更准确。

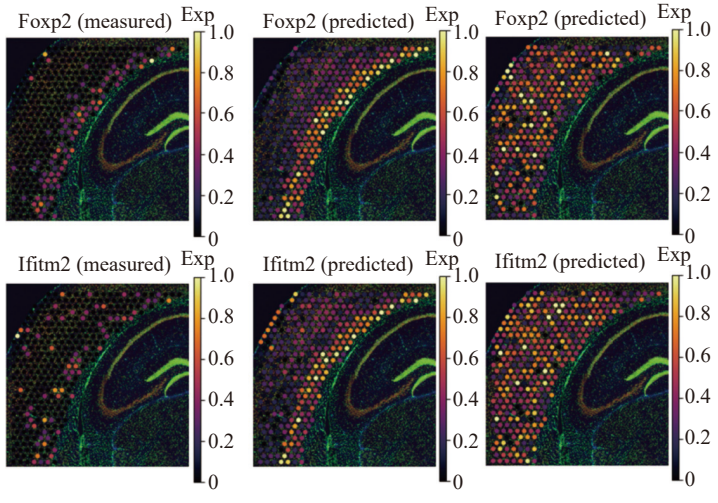
b. $\lambda_1 = \lambda_2 = 1$时，使用带有过滤项的损失函数进行映射，令$n_{\text{target}_{\text{cells}}} = n_{\text{seg}}$，$n_{\text{seg}}$为 Visium[2] 中细胞分割的细胞总数。细胞分割主要使用 squidpy[23] 包对图像进行分割，并统计每个网格点内细胞分割数量和坐标等信息。与 Tangram 相比，RK-Tangram 的余弦相似度更高(图 3(a))，预测的更精准。RK-Tangram 纠正了 Visium[2] 中测量的低质量基因、基因 Foxp2[24] 和 Ifitm2 在神经元 L6 CT 与非神经元 VLMC 细胞中的特异性表达(图 3(b))，而空间测序 Visium 检测到的表达较稀疏(图 3(c)左)。RK-Tangram 预测的结果更符合实际的基因表达水平(图 3(c)中)，细胞类型分类更加明显，而 Tangram 预测的结果不具有特异性(图 3(c)右)。因为 Visium[2] 为非靶向空间技术，一个空间体素网格中可能包含多个细胞，因此将其分解为单细胞是有必要的。通过映射矩阵得到每个单细胞所属体素网格点以及每个网格内的单细胞数，由于单细胞数据包含细胞类型注释，于是可以得



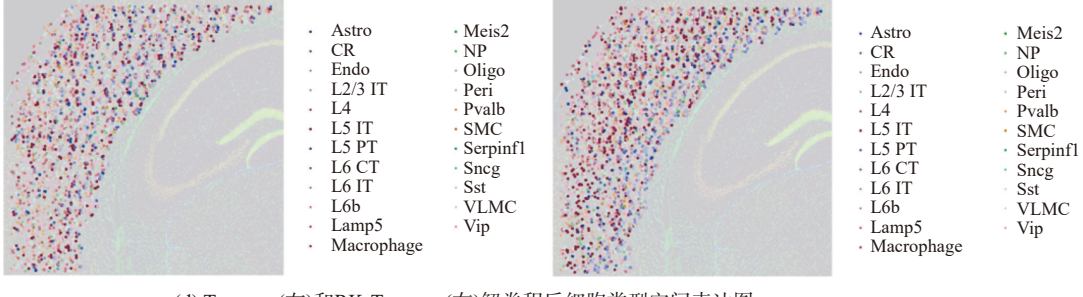
(a) 加上过滤项余弦相似度对比图



(b) 基因Foxp2(左)和Ifitm2 (右)表达UMAP图



(c) 原始空间测序(左)、RK-Tangram(中)与Tangram(右)预测基因空间表达图



(d) Tangram(左)和RK-Tangram(右)解卷积后细胞类型空间表达图

图 3　成年小鼠大脑皮质层数据带过滤项分析

Fig.3　Analysis of adult mouse brain cortical layer data with filtered items

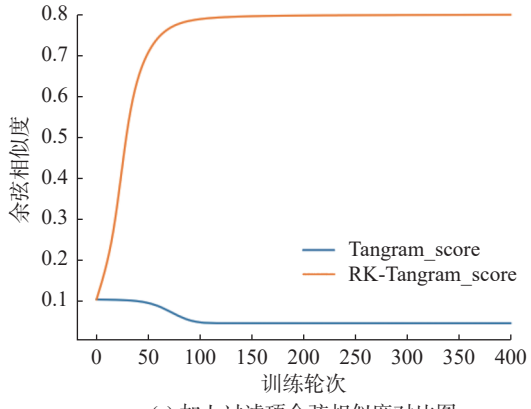
到每个网格点内不同细胞类型数。最后给每个体素内特定分割细胞随机分配细胞类型注释，并可视化分割细胞的细胞类型空间表达。由图 3(d)可知，RK-Tangram 比 Tangram 预测的细胞类型分层

更加明显，分层效果更好。

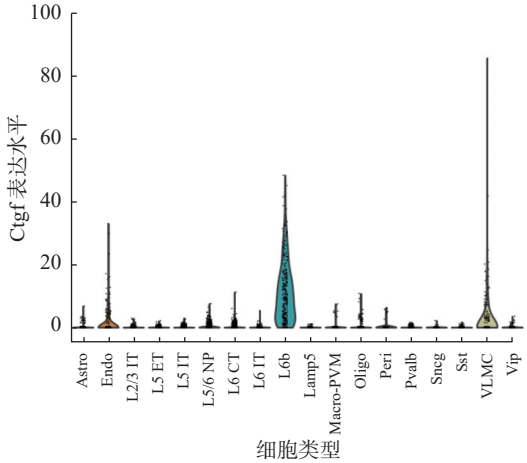
2.2.2　健康成年小鼠脑初级运动区（MOp）snRNAseq 单核数据与 slideseq 空间数据

第二组数据集测量的是健康成年小鼠大脑初级运动区域，包含单核 snRNAseq 测序数据和以高空间分辨率来测量全基因组表达的可扩展技术 Slideseq[3] 数据。单核 snRNAseq 数据包含 26431 个细胞上测量的 27742 个基因[18]；空间 Slideseq[3] 数据包含 9852 空间体素上测量的 24518 个基因[18]。对单核数据中每个细胞内的计数数据进行归一化处理，提供 253 个基因作为标记基因，经过筛选过滤，保留两个数据集共有的 249 个基因作为训练基因。

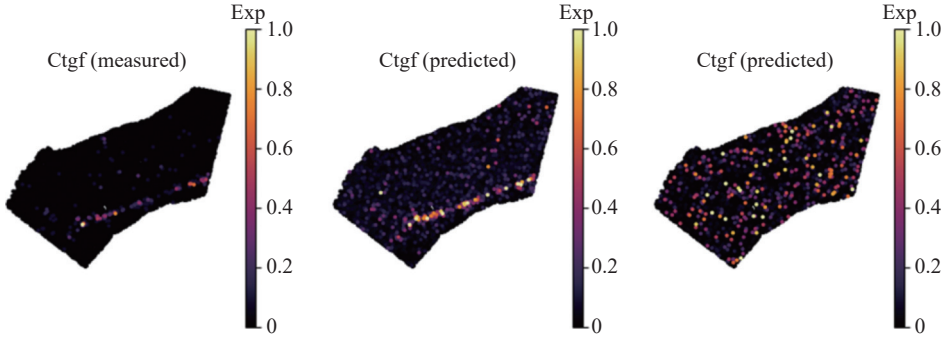
由余弦相似度图 4(a)可知，RK-Tangram 预测的空间表达矩阵与实际的 Slideseq 数据更接近；由基因 Ctgf 小提琴图 4(b)可知，其在谷氨酸能神经元亚类 L6b 区域中具有特异性表达[25]，RK-Tangram 预测特异性基因 Ctgf 的表达效果比 Tangram 好(图 4(c))，更符合实际的生物学意义。同时，基



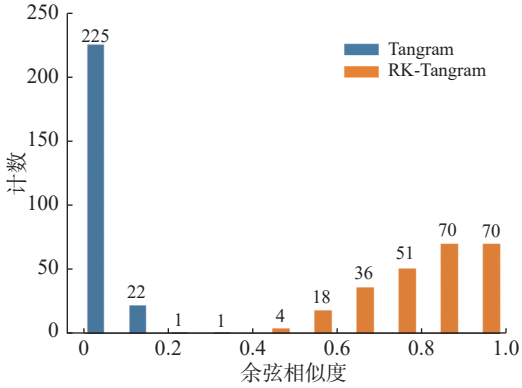
(a) 加上过滤项余弦相似度对比图



(b) 基因 Ctgf 的小提琴图



(c) 原始空间测序(左)、RK-Tangram(中)与Tangram(右)预测基因空间表达图



(d) 训练基因得分统计图

图 4　健康成年小鼠脑的初级运动区（MOp）数据分析

Fig.4　Analysis of primary motor area (MOp) data in the brain of healthy adult mouse

因 Ctgf 是测试集里的基因，这也说明 RK-Tangram 的泛化能力较强。另外通过训练基因得分统计图 4(d)可知，RK-Tangram 预测的基因余弦相似度大部分都在 0.5 以上，而 Tangram 预测的都在 0.5 以下，进一步说明 RK-Tangram 预测效果较好。

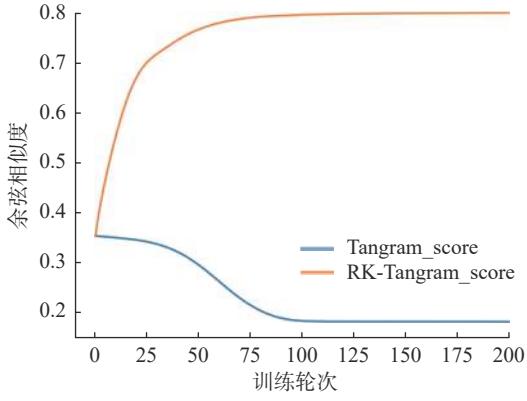
2.2.3 小鼠脑切片视觉区(VISp)的 snRNAseq 单核数据与 MERFISH 空间数据

snRNAseq 单核数据包含 11 759 个单细胞和 40 056 个基因的表达值，MERFISH 技术虽然提高了分辨率，但在单细胞水平上实现空间分辨的高度多重化 RNA 分析时，基因测量数量较少。此次使用的空间 MERFISH[6-7] 数据包含 2 399 个空间网格和 268 个基因的表达值[18]。这里提供 1 386 个基因作为 VISp 区的标记基因，筛选出 256 个基因作为训练基因进行深度学习。

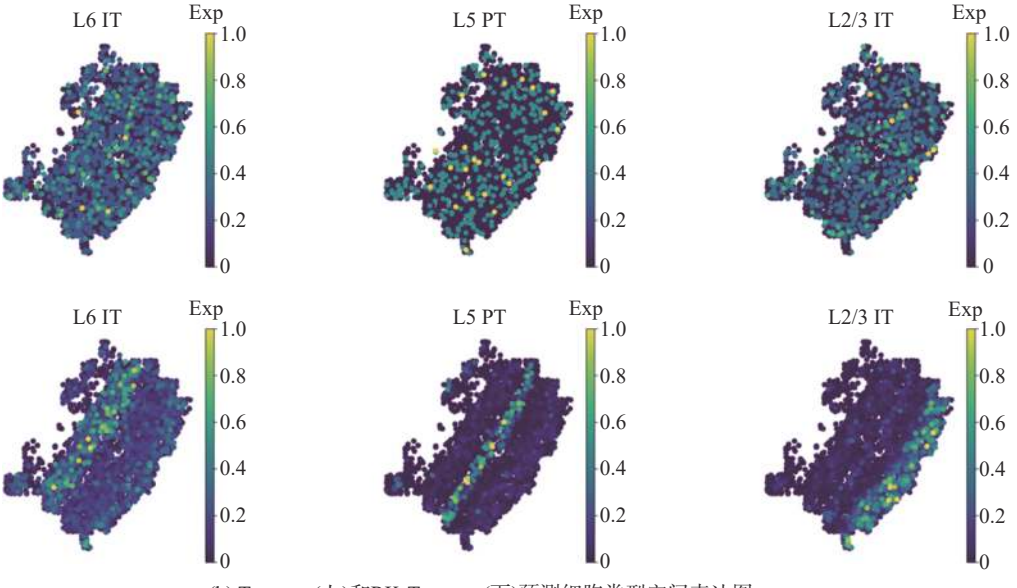
由余弦相似度曲线图 5(a)可知，与 Tangram 相比，RK-Tangram 预测的结果更准确。从细胞类型 L6 IT，L5 PT 和 L2/3 IT 的空间表达图 5(b)来看，RK-Tangram 预测的更具有聚类效果，而 Tangram 预测的却没有明显的细胞类型分层。

2.3 讨 论

在处理高通量数据时，Tangram[18] 损失函数值集中在熵项，这使得 Tangram 余弦相似度随着迭代步数的增加而递减，如图 1(d)所示。本文通过建立新的损失函数，并使用改进的 Adam 更新映射矩阵，使 RK-Tangram 损失函数中的余弦相似度灵敏度更高，从而避免了余弦相似度随着迭代次数增加而递减的问题，使预测效果更加显著。从模拟数据和真实数据分析来看，RK-Tangram 能够更好地通过单细胞数据预测空间数据，且对于测试集的泛化能力也更强，同时为空间数据分配的细胞类型与实际更加吻合，对于生物学研究更有意义。



(a) 余弦相似度对比图



(b) Tangram(上)和RK-Tangram(下)预测细胞类型空间表达图

图 5 小鼠脑切片视觉区(VISp)数据分析

Fig.5 Analysis of visual areas (VISp) in mouse brain slices

3　结　论

通过巧妙的构造损失函数和修正梯度值，本文设计了 RK-Tangram 算法。为了评估算法性能，将算法应用到 3 组模拟数据和 3 组真实的小鼠大脑数据。对比实验结果表明，相比于 Tangram，RK-Tangram 具有更高的余弦相似度、更快的收敛速度和更小的损失值，同时 RK-Tangram 预测的空间表达数据更符合实际生物学特性，对生物学和病理学研究具有重要的意义。

另外，优化算法 Adam 为一阶梯度下降法的推广，而牛顿法使用二阶泰勒展开计算，精度比 Adam 高，但目前 Hessian 矩阵的计算面临一定的困难，如果能解决这一问题，利用牛顿法计算映射矩阵，可能会进一步提高预测精度。

参考文献：

[1] JOVIC D, LIANG X, ZENG H, et al. Single-cell RNA sequencing technologies and applications: A brief overview[J]. Clinical and Translational Medicine, 2022, 12(3): e694.

[2] STÅHL P L, SALMÉN F, VICKOVIC S, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics[J]. Science, 2016, 353(6294): 78−82.

[3] RODRIQUES S G, STICKELS R R, GOEVA A, et al. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution[J]. Science, 2019, 363(6434): 1463−1467.

[4] STICKELS R R, MURRAY E, KUMAR P, et al. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2[J]. Nature Biotechnology, 2021, 39(3): 313−319.

[5] VICKOVIC S, ERASLAN G, SALMÉN F, et al. High-definition spatial transcriptomics for in situ tissue profiling[J]. Nature Methods, 2019, 16(10): 987−990.

[6] MOFFITT J R, BAMBAH-MUKKU D, EICHHORN S W, et al. Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region[J]. Science, 2018, 362(6416): eaau5324.

[7] CHEN K H, BOETTIGER A N, MOFFITT J R, et al. Spatially resolved, highly multiplexed RNA profiling in single cells[J]. Science, 2015, 348(6233): eaaa6090.

[8] CHEN J X, MCSWIGGEN D, ÜNAL E. Single molecule fluorescence *in situ* hybridization (smFISH) analysis in budding yeast vegetative growth and meiosis[J]. Journal of Visualized Experiments, 2018, (135): 57774.

[9] CODELUPPI S, BORM L E, ZEISEL A, et al. Spatial organization of the somatosensory cortex revealed by osmFISH[J]. Nature Methods, 2018, 15(11): 932−935.

[10] WANG X, ALLEN W E, WRIGHT M A, et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states[J]. Science, 2018, 361(6400): eaat5691.

[11] LUBECK E, COSKUN A F, ZHIYENTAYEV T, et al. Single-cell *in situ* RNA profiling by sequential hybridization[J]. Nature Methods, 2014, 11(4): 360−361.

[12] SHAH S, LUBECK E, ZHOU W, et al. In situ transcription profiling of single cells reveals spatial organization of cells in the mouse hippocampus[J]. Neuron, 2016, 92(2): 342−357.

[13] ENG C H L, LAWSON M, ZHU Q, et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+[J]. Nature, 2019, 568(7751): 235−239.

[14] LÄHNEMANN D, KÖSTER J, SZCZUREK E, et al. Eleven grand challenges in single-cell data science[J]. Genome Biology, 2020, 21(1): 31.

[15] ZHAO E, STONE M R, REN X, et al. Spatial transcriptomics at subspot resolution with BayesSpace[J]. Nat Biotechnol, 2021, 39(11): 1375-1384.

[16] KLESHCHEVNIKOV V, SHMATKO A, DANN E, et al. Comprehensive mapping of tissue cell architecture via integrated single cell and spatial transcriptomics[EB/OL]. (2020-11-17). https://www.biorxiv.org/content/10.1101/2020.11.15.378125v1.abstract

[17] ELOSUA-BAYES M, NIETO P, MEREU E, et al. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes[J]. Nucleic Acids Research, 2021, 49(9): e50.

[18] BIANCALANI T, SCALIA G, BUFFONI L, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram[J]. Nature Methods, 2021, 18(11): 1352−1362.

[19] KINGMA D P, BA J. Adam: A method for stochastic optimization[C]//3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

[20] HE Y L, ZHANG X L, AO W, et al. Determining the optimal temperature parameter for Softmax function in reinforcement learning[J]. Applied Soft Computing, 2018, 70: 80−85.

[21] HUYBRECHS D, TREFETHEN L N. Sigmoid functions and multiscale resolution of singularities[EB/OL]. (2023-03-03). https://arxiv.org/abs/2303.01967

[22] HARRIS C R, MILLMAN K J, VAN DER WALT S J, et al. Array programming with NumPy[J]. Nature, 2020, 585(7825): 357−362.

[23] PALLA G, SPITZER H, KLEIN M, et al. Squidpy: a scalable framework for spatial omics analysis[J]. Nature Methods, 2022, 19(2): 171−178.

[24] TASIC B, YAO Z Z, GRAYBUCK L T, et al. Shared and distinct transcriptomic cell types across neocortical areas[J]. Nature, 2018, 563(7729): 72−78.

[25] ZOLNIK T A, LEDDEROSE J, TOUMAZOU M, et al. Layer 6b is driven by intracortical long-range projection neurons[J]. Cell Reports, 2020, 30(10): 3492–3505.

（编辑：丁红艺）