DOI:10.12113/202311005

基于迁移学习的空间转录组数据解卷积算法

陈子睿[1],杨博然[1],何田韵[1],李 建[2*],周光华[3*]
(1. 哈尔滨工业大学 数学学院,哈尔滨 150001;2. 成都医学院,成都 610500;
3. 国家卫生健康委统计信息中心, 北京 100044)

摘 要:空间转录组(Spatial transcriptomics, ST)测序技术可以捕获多个细胞的空间位置信息,但无法达到单细胞分辨率,阻碍了对细胞类型异质性空间模式和基因表达特异性的解析。针对ST数据,本文提出了基于DenseNet网络结构和CORAL域自适应理论的细胞类型解卷积算法(STDN)。STDN通过学习引入的单细胞转录组(Single cell RNA sequencing, scRNA-seq)数据的细胞类型信息,利用迁移学习模型将其迁移到ST数据上,从而达到预测ST数据中每个捕获位点(Spot)的细胞类型组成及比例的目的。本文通过4组scRNA-seq真实数据及模拟的配套ST数据,表明STDN可以有效地恢复细胞类型转录谱及其在Spots内的比例,且优于其它解卷积算法。STDN对小鼠海马体和人类胰腺导管腺癌的ST数据进行解卷积,确定了组织中的多种细胞类型,解析了组织和癌症的高度异质性,为研究疾病的致病机理奠定了基础。

关键词:空间转录组数据;解卷积;DenseNet;CORAL
中图分类号:O213; **文献标志码**:A **文章编号**:1672-5565(2025)01-040-10

A deconvolution algorithm based on transfer learning for spatial transcriptomics

CHEN Zirui[1], YANG Boran[1], HE Tianyun[1], LI Jian[2*], ZHOU Guanghua[3*]
(1.*School of Mathematics*, *Harbin Institute of Technology*, *Harbin* 150001, *China*;2. *Chengdu Medical College*, *Chengdu* 610500, *China*;
3. *Center of Statistics and Health Informatics*, *National Health Commission of People's Republic of China*, *Beijing* 100044, *China*)

Abstract:Spatial transcriptomics sequencing technology captures spatial location information of multiple cells, but single-cell resolution cannot be achieved, which hampers the analysis of spatial patterns of cell type heterogeneity and gene expression specificity. The cell type deconvolution algorithm (STDN) based on DenseNet network structure and CORAL domain adaptive theory is proposed for spatial transcriptomics data. STDN learns cell type information about introduced single-cell RNA sequencing (scRNA-seq) data and migrates it to ST data using a transfer learning model. Thus, the purpose of predicting the cell type composition and proportion of each capture site (Spot) in the ST data is achieved. In this paper, four factual scRNA-seq datasets and simulated matching ST datasets show that STDN can effectively recover cell type transcription profiles and their proportions in Spots, and is superior to other deconvolution algorithms. By deconvolution of ST data from mouse hippocampus and human pancreatic ductal adenocarcinoma, STDN identifies multiple cell types in tissues, resolves the high heterogeneity of tissues and cancers, and laid a foundation for studying the pathogenesis of the disease.

Keywords:Spatial transcriptomics data; Deconvolution; DenseNet; CORAL

收稿日期:2023-11-07;**修回日期**:2024-01-10;**网络首发日期**:2024-04-15.
网络首发地址:https://link.cnki.net/urlid/23.1513.Q.20240412.2122.002
*通信作者:李建,研究实习员,研究方向:生物信息学,E-mail:ianlly@163.com;
周光华,副研究员,研究方向:生物信息学,E-mail:zhough1101@163.com.
引用格式:陈子睿,杨博然,何田韵,等.基于迁移学习的空间转录组数据解卷积算法[J].生物信息学,2025,23(1):40-49.
CHEN Zirui, YANG Boran, HE Tianyun, et al. A deconvolution algorithm based on transfer learning for spatial transcriptomics[J].Chinese Journal of Bioinformatics,2025,23(1):40-49.

描绘组织或疾病的不同细胞类型的空间结构对于解析组织或疾病中细胞功能和分子结构至关重要[1]。空间转录组测序技术的出现,使得从空间环境中研究组织的基因表达谱成为可能[2]。空间转录组测序技术对每个捕获位点(Spot)进行测序,在获取基因表达信息的同时保留了空间位置信息,然而,每个 Spot 包含多个细胞,测量值则为潜在异质细胞类型细胞混合物的平均基因表达。单细胞测序技术在将实体组织解离成单个细胞进行测序的过程中,虽然丢失了空间位置信息,但可得到单个细胞的转录本信息[3]。因此,利用空间和基因组信息解析每个 Spot 中的细胞类型及其占比,即对 ST 数据进行细胞类型反卷积,成为了解析细胞类型的空间位置和表征复杂组织结构的关键步骤[4]。

最近大量针对 ST 数据的细胞类型解卷积算法被开发,大致分为基于统计学习和机器学习两类算法。SPOTlight[5]采用非负矩阵分解回归作为核心算法,考虑先验信息,借助细胞类型的 Marker 基因初始化基矩阵与系数矩阵,使用非负最小二乘法(Non-negative least squares)计算 Spot 的系数矩阵及其细胞类型组成,从推断复杂组织内细胞类型及状态的空间分布。SpatialDWLS[6]通过细胞类型富集分析确定可能存在于每个 Spot 的细胞类型,利用阻尼加权最小二乘法(Damped least squares)选择使总体相对错误率最小的权重,从而定量估计每个 Spot 的细胞类型的精确组成。RCTD[7]以带有细胞类型注释的 scRNA-seq 数据作为参考,在消除样本批次效应的同时识别 ST 数据每个 Spot 的细胞类型,再通过拟合统计模型精确解卷积。Cell2location[8]是基于贝叶斯模型构建的,考虑了数据样本的不同技术来源,借用了跨位置的统计强度,从而实现单细胞测序数据和空间转录组数据的整合。该算法可解析空间转录组测序数据中的细粒度细胞类型,并具有创建不同组织的综合细胞图谱的功能。CARD[9]是一种基于条件自回归的解卷积方法,它建立在非负矩阵分解模型的基础上,基于条件自回归建模假设,综合考虑 scRNA-seq 数据的细胞类型特异性基因表达信息和空间相关性对 ST 数据进行卷积。Stereoscope[10]基于单细胞转录组和空间转录组数据均遵循负二项分布(Negative binomial distribution)的假设,使用概率模型描述单细胞转录组和空间转录组数据之间的关系,对空间转录组的细胞类型混合物进行引导解卷积,从而将单细胞转录组细胞类型映射到空间转录组。

近些年,机器学习被广泛用于空间转录组数据解卷积算法的研究。DSTG[11]使用共享最近邻学习 ST 模拟数据和 ST 真实数据之间的链接图,以获得的链接图为基础,通过半监督图卷积网络学习局部图结构和基因表达模式的潜在表示,进而预测 ST 真实数据中的细胞类型组成。STRIDE[12]对带细胞类型标签注释的 scRNA-seq 数据进行主题建模,根据在 scRNA-seq 数据中获得的基因-主题分布,使用隐狄利克雷分布(Latent dirichlet allocation, LDA)估计 Spot-主题分布,而后整合 Spot-主题分布和细胞类型-主题分布计算各 Spots 包含每种细胞类型的概率,将这种概率视为每种细胞类型在该 Spot 中所占的比例,从而达到对 ST 数据进行细胞类型分解的目的。DestVI[13]是一种用于 ST 数据中细胞类型的多分辨率解卷积的贝叶斯模型。该算法引入了解码器神经网络的变分推理模型,它使用条件深度生成模型学习离散的细胞类型特异性分布和连续亚细胞类型潜在变化。在 Spots 中的转录本数量遵循负二项分布的前提假设下,DestVI 分别为 scRNA-seq 数据(scLVM)和 ST 数据(stLVM)构建了不同的潜在变量模型(LVM). stLVM 使用由 scLVM 训练的解码器神经网络,并使用最大后验(MAP)估算细胞类型比例。张柳[14]构建了一种基于统计学习方法和迁移学习的方法,该模型将原本用于图片识别的子空间对齐领域自适应迁移学习方法用于 scRNA-seq 和 ST 数据,通过对齐源域和目标域的子空间实现 ST 数据的降维,批次效应校正和细胞类型预测。CellDART[15]从单细胞转录组数据中随机选择细胞构成一个细胞类型比例已知的伪 Spot,从伪 Spot 的基因表达中提取细胞部分信息训练神经网络模型,应用于空间转录组数据的不同 Spot,有望帮助阐明细胞的空间异质性及其在各种组织中的紧密相互作用。

尽管存在多种 ST 数据解卷积算法,但这些算法仍存在数据信息提取不够充分、对数据敏感性高等问题。例如,基于统计学习方法的 SPOTlight 和 RCTD 没有将捕获的位置信息合并到模型空间分解中,SpatialDWLS 估计稀有细胞类型的比例时偏差较大;基于机器学习方法的 DSTG 高度依赖建模图卷积神经网络的链接图的质量,CellDART 特异性不够高等。因此,迫切需要研究新方法,深度整合单细胞和空间转录组数据中的互补信息,实现多源生物数据的整合,从而更准确推断每个组织 Spot 的细胞组成。

本文开发了一种基于深度迁移学习模型的细胞类型解卷积算法 STDN,用于解析 ST 数据中每个 Spot 的细胞类型组成。STDN 是基于 DenseNet 和相关对齐(CORrelation alignment, CORAL)[16]搭建的,

通过共享基因集建立 scRNA-seq 数据集和 ST 数据集之间的联系,再引入迁移学习中的领域自适应技术来充分挖掘两类数据集潜在的信息特征,从而利用从 scRNA-seq 数据中学到的细胞类型知识来解决 ST 数据中每个 Spot 的细胞类型识别问题,更精确地实现 ST 数据的细胞类型解卷积。

1 方法

1.1 STDN 算法

为了整合 scRNA-seq 和 ST 数据,以实现 ST 数据解卷积,本文提出了 STDN 算法,图 1 的三幅子图分别展示了 STDN 建模的整体流程(a)、STDN 模型训练过程(b)和 STDN 模型预测过程(c)。

STDN 的输入为 scRNA-seq 和 ST 数据,输出为 ST 数据中每一个 Spot 包含的细胞类型及其占比。首先,对 scRNA-seq 和 ST 数据进行共享特征选择。然后,STDN 基于 DenseNet 模型,使用基因表达数据和细胞类型标签进行模型训练,在隐藏层使用 CORAL 减少单细胞转录组和空间转录组之间的差异,学习提取与细胞类型相关的特征用于模型预测。最后,使用训练好的模型对空间转录组数据进行预测,将单细胞转录组数据的细胞类型标签转移到 Spot 上,从而对每个 Spot 中的细胞类型及其占比进行估计。

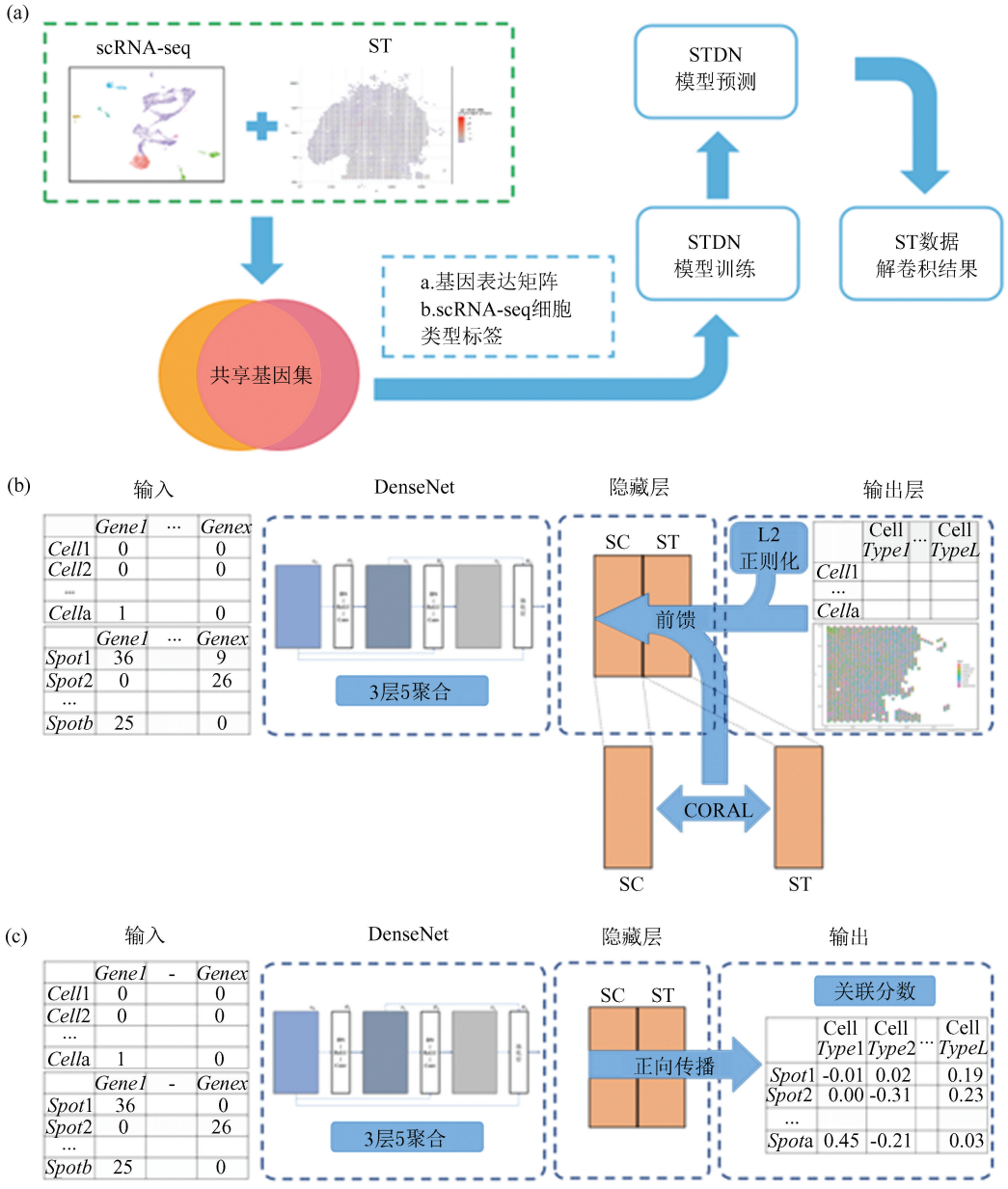

图 1 (a) STDN 流程图;(b) STDN 模型训练 ;(c) STDN 模型预测

Fig.1 (a) Pipeline of STDN;(b) STDN model training;(c) STDN model prediction

1.2　共享特征选择

特征选择是机器学习中必要的数据预处理步骤,旨在从已有的特征集合中选择与任务相关的特征子集以提高模型运算效率,同时确保不丢失重要信息。生物基因测序数据具有高维度的特点,因此在分析之前尤其需要进行特征选择。生物大数据的分析过程中容易遇到特征过多造成的维数灾难问题,与此同时,去除不相关特征还可以排除无关元素的干扰,降低学习任务的难度,提高任务解决的效率。

除此之外,特征基因需在 scRNA-seq 数据和 ST 数据中均存在,则在模型训练之前对 scRNA-seq 和 ST 数据的基因取交集,从而得到共享特征空间,称为“共享基因集”。

1.3　STDN 模型训练

STDN 引入了一种深度迁移学习模型,迁移学习是从源域获取解决问题的知识并储存,然后应用该知识来解决目标域中类似问题的一种机器学习方法。

STDN 模型以处理好的 scRNA-seq 数据和 ST 数据的基因表达矩阵以及 scRNA-seq 数据的细胞类型 one-hot 矩阵为输入,基因表达矩阵行为细胞或 Spot,列为基因。经过 3 层 5 次聚合的 DenseNet 网络训练,选择 Sigmoid 激活函数,同时以细胞类型分类错误损失、度量 scRNA-seq 和 ST 基因表达之间潜在差异的 CORAL 损失和防止过拟合的 L_2 正则化损失 3 种损失函数构成加权损失函数,损失函数在各层间反向传播信息,训练模型。具体模型训练流程如图 1(b)所示。

STDN 使用 CORAL 度量源域和目标域中数据分布差异,这是一种非对称转换,它计算的是源域和目标域特征的二阶统计量(协方差)之间的距离。设源域训练样本为 $D_S = \{x_i\}$, $x_i \in \mathbb{R}^d$,标签为 $L_S = \{y_i\}$, $i \in \{1,\cdots,L\}$. 目标域数据为 $D_T = \{u_i\}$, $u_i \in \mathbb{R}^d$, 其中 d 为网络的输出个数,即神经元个数,目标域无标签。CORAL 损失计算公式如下:

$$Loss_{CORAL} = \frac{1}{4d^2} \| C_S - C_T \|_F^2 \tag{1}$$

其中 $\| \cdot \|_F^2$ 表示 Frobenius 范数, C_S 和 C_T 分别表示源域和目标域的特征协方差矩阵,计算公式如下:

$$C_S = \frac{1}{n_S - 1}\left(D_S^{\mathrm{T}} D_S - \frac{1}{n_S}(1^{\mathrm{T}} D_S)^{\mathrm{T}}(1^{\mathrm{T}} D_S)\right) \tag{2}$$

$$C_T = \frac{1}{n_T - 1}\left(D_T^{\mathrm{T}} D_T - \frac{1}{n_T}(1^{\mathrm{T}} D_T)^{T}(1^{\mathrm{T}} D_T)\right) \tag{3}$$

其中 n_S 和 n_T 分别表示源域样本和目标域样本个数, D_S^{ij} 和 D_T^{ij} 分别表示源域和目标域第 i 个样本的第 j 个特征。

STDN 模型使用了 3 层 DenseNet 网络,训练了 5 个模型进行聚合。下面以单层网络为例介绍 STDN 模型设置:首先,使用一个共同的隐藏层来合并 scRNA-seq 和 ST 的基因表达矩阵,隐藏层激活函数设置如下:

1)利用隐藏层使用 Sigmoid 激活函数将基因表达矩阵转化至较低维度,隐藏层激活函数设置为

$$f_{\mathrm{Hidden}}(X) = \mathrm{sigmoid}(X^{\mathrm{T}}\theta_{\mathrm{Hidden}} + b_{\mathrm{Hidden}}) \tag{4}$$

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{5}$$

其中, X 表示基因表达矩阵, θ_{Hidden} 表示隐藏层权重, b_{Hidden} 表示隐藏层偏差。

2)添加细胞类型输出层。空间转录组学数据可以有分类输出或无输出,无输出意味着没有匹配的细胞类型。将细胞类型层激活函数设置为

$$f_{\mathrm{Class}}(X) = \mathrm{sigmoid}(f_{\mathrm{Hidden}}(X)^{\mathrm{T}}\theta_{\mathrm{Class}} + b_{\mathrm{Class}}) \tag{6}$$

其中, X 表示基因表达矩阵, θ_{Class} 表示细胞类型层权重, b_{Class} 表示细胞类型层偏差。

接下来,添加输出层以预测空间转录组数据各 Spot 的细胞类型及比例。为了训练模型,需要分别计算分类输出和 CORAL 两种损失函数,公式如下:

$$Loss_{\mathrm{Class}} = \frac{1}{n}\sum_{i=1}^{n}(Y_{\mathrm{Type},i} - f_{\mathrm{Class}}(X)_i) \tag{7}$$

$$Loss_{\mathrm{CORAL}} = \mathrm{CORAL}(X_{\mathrm{Cell}}, X_{\mathrm{Spot}}) \tag{8}$$

其中, $Loss_{\mathrm{Class}}$ 表示分类输出损失函数, $Loss_{\mathrm{CORAL}}$ 表示 CORAL 损失函数。 $Y_{\mathrm{Type},i}$ 表示真实细胞类型, X 表示输入的基因表达矩阵。 X_{Cell} , X_{Spot} 分别代表 scRNA-seq 数据基因表达矩阵和空间转录组数据基因表达矩阵,最小化 CORAL 损失是映射 scRNA-seq 数据和 ST 数据之间的数据代表性分布的关键。

另外,为了避免过拟合,模型添加一个 L_2 正则化损失。因此,该模型总体损失函数是上述三种损失函数的加权和,见(9)式。

$$\mathrm{Loss} = \lambda_1 \mathrm{Loss}_{\mathrm{Class}} + \lambda_2 \mathrm{Loss}_{\mathrm{CORAL}} + \lambda_3 \| \theta \|_2^2 \tag{9}$$

其中, λ_1 , λ_2 , λ_3 表示三种损失函数的权重,可以用来调整每个损失项和正则化项的重要性。损失函数反向传播达到模型训练的目的。

1.4　模型预测

训练好的迁移学习模型被应用于 ST 数据细胞类型及其占比预测,具体流程如图 1(c)所示。模型输出是 Sigmoid 激活函数的输出,先对输出结果进行 KNN 平滑处理以消除随机噪声,后将输出值域为[0,1]的结果转换为类似相关系数的关联得分,公

式如下：

$$AS = 2\left(\frac{\text{Preds\_Smooth} - \frac{1}{L}}{2 - \frac{2}{L}} + \frac{1}{2}\right) - 1 \quad (10)$$

其中 AS 表示关联得分(Associate score)，$Preds\_Smooth$ 表示平滑处理后的模型输出预测值，L 表示预测标签数目。该关联得分代表包含某种类型细胞的可能性，最终得到以 Spot 为行，以细胞类型为列的得分矩阵，反映 ST 数据每个 Spot 包含的某种细胞类型的可能性。

1.5 模型效果评估指标

本文使用 JS 散度（Jensen-shannon divergence，JSD）[17]进行模型预测性能评估，这是一种衡量两个概率分布之间差异的信息熵方法，它对 KL 散度(Kullback-leibler divergence)进行了改进，解决了 KL 散度的不对称问题。KL 散度及 JS 散度计算公式见(11)式、(12)式。

$$KL(P\|Q) = \sum P(x)\log_2\frac{P(x)}{Q(x)} \quad (11)$$

$$JS(P\|Q) = \frac{1}{2}KL\left(P(x)\,\|\,\frac{P(x)+Q(x)}{2}\right) + \frac{1}{2}KL\left(Q(x)\,\|\,\frac{P(x)+Q(x)}{2}\right) \quad (12)$$

其中，$P(x)$ 表示真实细胞类型的分布，$Q(x)$ 表示训练后的 STDN 模型预测的细胞类型分布。JS 散度的值域范围是[0,1]，值越小表示两种分布越接近。可以通过这种方式计算所有 Spots 的 JS 散度，然后将所有 Spots 的 JS 散度的平均值作为评估指标，称为 JSD 分数。JSD 分数越低说明预测细胞类型和真实细胞类型的分布越相似，则表明模型预测性能越好。

2 结果与讨论

2.1 基准测试

2.1.1 模拟数据集介绍

从文献[18]中收集了 4 组配套的 scRNA-seq 数据和 ST 模拟数据。4 组 scRNA-seq 数据中 2 组数据来自人类健康胰腺组织，另外 2 组分别来自小鼠健康胰腺组织和小鼠健康气管组织，配套的 ST 数据由 scRNA-seq 数据模拟生成。

参考 RCTD 和 Stereoscope 的空间转录组数据模拟生成方法，设计了配套空间转录组数据的模拟生成过程。对于每个模拟 Spot，首先采样细胞数目在 5~15 的均匀分布和采样细胞类型数目在 2~6 的均匀分布。然后假设这些细胞类型分布的可能性是相等的，并从 scRNA-seq 数据的每种细胞类型中随机分配细胞到该 Spot。为了获得每个 Spot 的基因表达值，将一个 Spot 上所有细胞的基因表达值进行求和作为该模拟 Spot 的基因表达值。参考 RCTD 中空间转录组模拟数据集的构造方法，使用 Scuttle 包（http://bioconductor.org/packages/release/bioc/html/scuttle.html）将每个 Spot 的 Count 计数降采样到原始值的 10%. 通过计算对应于细胞类型的细胞数来获得每个 Spot 上细胞类型的百分比，将每个模拟数据集的 Spots 数量设置为1 000。4 组配套的 scRNA-seq 数据和 ST 模拟数据详细信息见表 1。

表 1 scRNA-seq 数据和 ST 模拟数据信息

Table 1 scRNA-seq data and the information of ST simulated data （个）

数据集	物种	组织	scRNA-seq 数据		ST 数据	
			细胞数	基因数	捕获位点数	基因数
数据集 1	人类	胰腺	1 040	21 572	1 000	17 499
数据集 2	人类	胰腺	943	21 413	1 000	21 198
数据集 3	小鼠	胰腺	1 382	19 479	1 000	14 860
数据集 4	小鼠	气管	6 937	27 083	1 000	18 388

2.1.2 ST 模拟数据的解卷积结果

为了可比较性，本文为不同数据集设置统一的超参数：训练步骤为 2 000，空间转录组数据一次训练所选取的样本数(Batch size)为 200，单细胞数据一次训练所选取的样本数(Batch size)为 50，隐藏层节点数为 50，Dropout 率为 50%，3 种损失函数的权重均设置为 3。

将模型预测结果与 ST 模拟数据集的基本事实进行比较，从而评估模型解卷积性能。将空间转录组模拟数据的真实细胞类型分布和模型预测的细胞类型分布进行相关性分析，得到真实-预测细胞类型相关矩阵，可视化结果如图 2 所示，图中展示了 4 组数据真实细胞类型和预测细胞类型概率分布矩阵的相关性。观察 4 组数据相关性图发现，相关性图中对角线位置相关系数更高，说明真实细胞类型和预测细胞类型相同的关联性更高，不同细胞类型相关性低，表明模型预测结果和基本事实吻合度较高，说明了 STDN 具有较高的准确性。

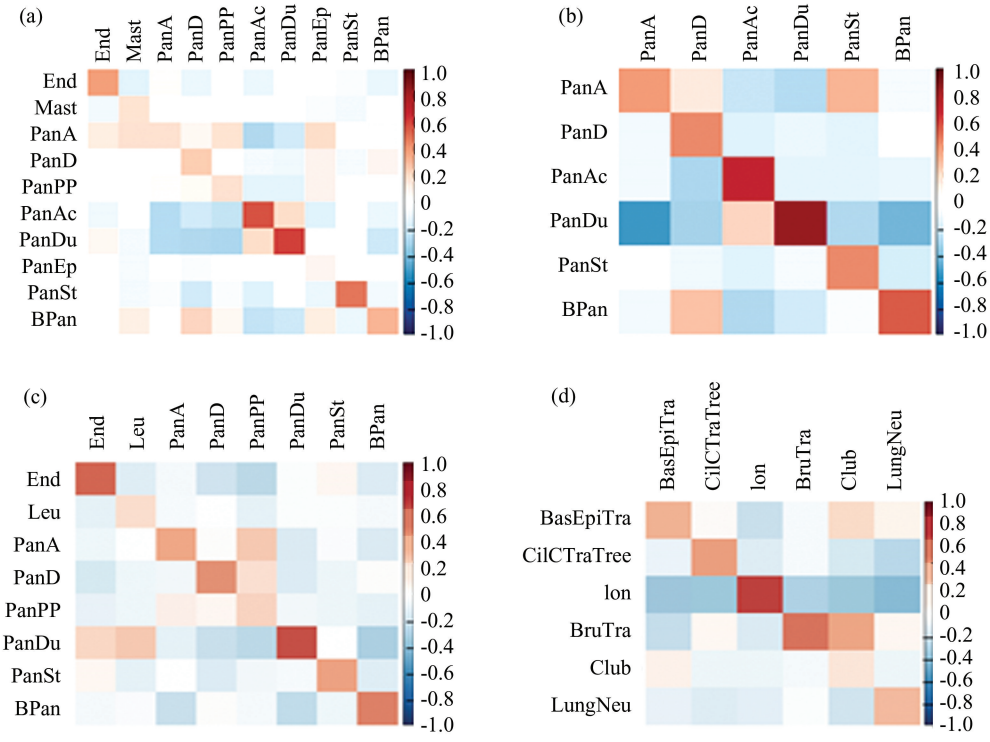

图 2　真实-预测细胞类型相关性图

Fig.2　Heatmap of ground truth and predicted value

注:(a) ST 模拟数据集 1;(b) ST 模拟数据集 2;(c) ST 模拟数据集 3;(d) ST 模拟数据集 4.

为了更好地测试本算法的预测性能,本文将 4 组模拟的 ST 数据已知细胞类型和 STDN 算法预测的细胞类型进行对比,用 JSD 指标度量预测细胞类型分布和真实细胞类型分布之间的差异。通过计算得到的所有 Spots 的 JS 散度,以这些 JS 散度的平均值为指标评估模型解卷积性能。本文利用 JSD 指标,将 STDN 算法与五种经典解卷积算法(RCTD, Seurat, SpatialDWLS, SPOTlight 和 DSTG)进行比较。最终得到的指标值如图 3,直观地展示了各算法解卷积性能的优劣。结果表明,相比于其他算法,STDN 在 4 组 ST 模拟数据集中表现较好,JS 散度指标值较低,分别为 0.41, 0.25, 0.37, 0.36, 说明 STDN 性能基本优于其他算法。

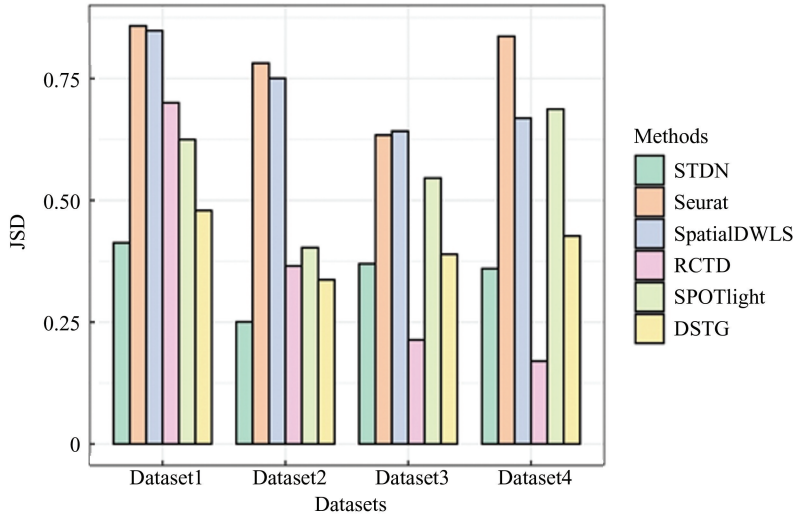

图 3　模型预测性能比较

Fig.3　Prediction performance on different methods

2.2 小鼠大脑海马体数据分析

2.2.1 小鼠大脑海马体数据集

从文献[19]和 GEO 数据库中收集了配对的出生后 7 d 小鼠大脑海马体 scRNA-seq 数据和 ST 数据,GEO 数据序列号为 GSM4800800, GSM4800808。

2.2.2 实验结果分析

为了更直观地展示解卷积结果,本文将细胞类型预测关联得分投影到 ST 数据组织切片的空间位置,结果如图 4 所示,包括出生后 7 d 小鼠大脑海马体空间转录组测序的组织切片图像(图 4(a))和 7 种细胞类型的预测结果(图 4(b)~4(g))。图 4(b)~4(g)分别反映了星形胶质细胞、内皮细胞、上皮细胞、成纤维细胞、小胶质细胞、神经元细胞、少突胶质细胞这 7 种类型细胞被每个 Spot 包含的可能性。每幅图横轴表示 ST 数据空间位置横坐标,纵轴表示 ST 数据空间位置纵坐标。图中一个点对应一个 Spot,颜色表示关联得分的大小,越接近红色表示与对应细胞类型关联性越强,即该 Spot 含有该类型细胞的可能性越大,越接近紫色表示该 Spot 含有该类型细胞的可能性越小。

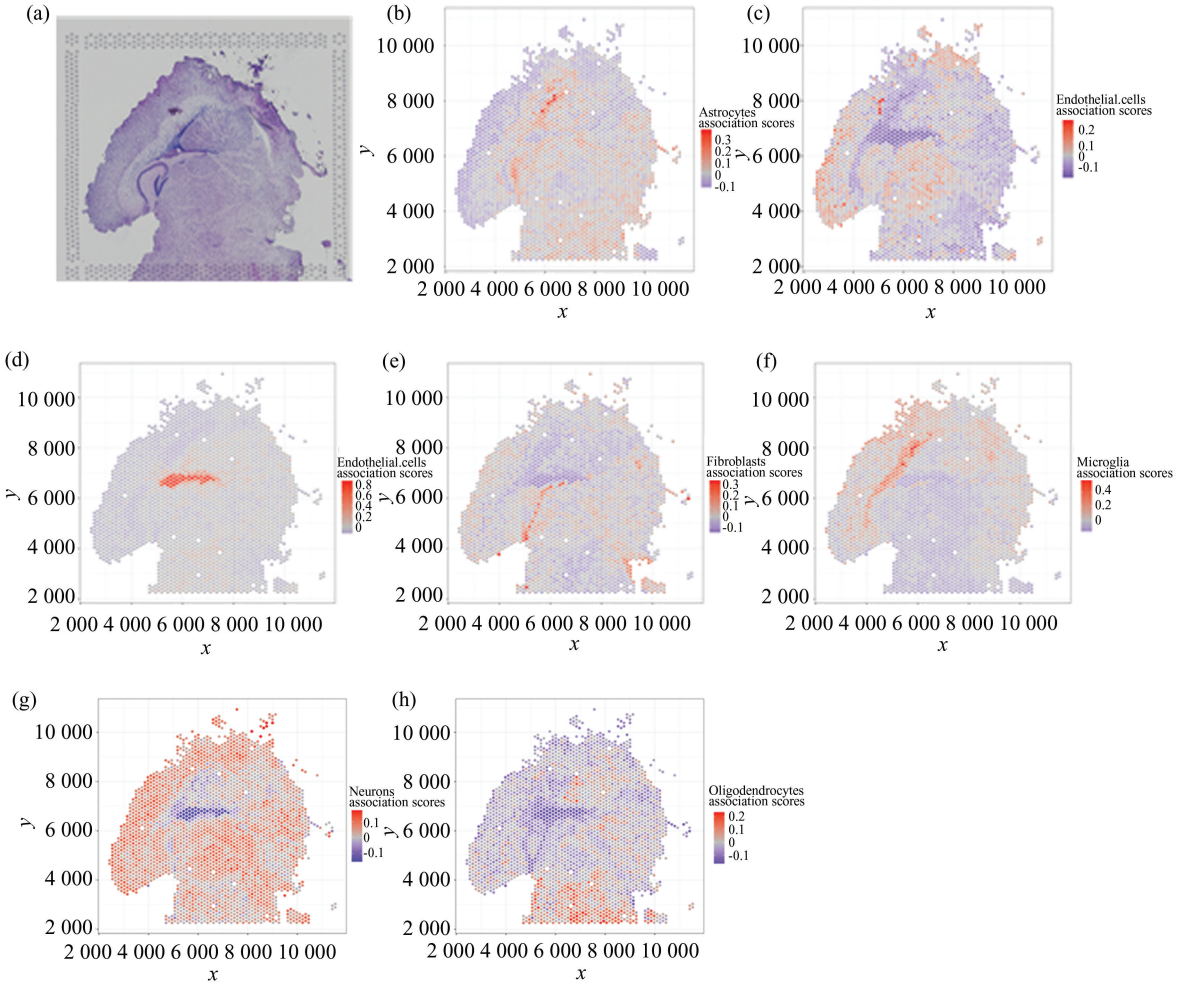

图 4 出生后 7 d 小鼠大脑海马体 ST 数据解卷积关联得分图

Fig.4 Deconvolution correlation score map of mouse brain hippocampus ST data 7 days after birth

注:图中 x,y 均代表 Spot 在组织切片中的空间位置坐标;(a) 组织切片图;(b) 星形胶质细胞关联得分;(c) 内皮细胞关联得分;(d) 上皮细胞关联得分;(e) 成纤维细胞关联得分;(f) 小胶质细胞关联得分;(g) 神经元细胞关联得分;(h) 少突胶质细胞关联得分.(扫本文首页二维码见彩图).

由图 4 可以明显地观察到细胞类型预测情况与组织切片的染色情况在一定程度上是一致的。例如,图 4(b)表示的星形胶质细胞高关联得分部位对应组织切片图的深紫色部位,图 4(d)表示的上皮细胞高关联得分部位对应组织切片图偏左上位置水平分布的浅紫色部位,图 4(e)表示的成纤维细胞高关联得分部位对应组织切片图偏左下角凹陷处的浅色区域。已有研究表明组织切片着色情况与细胞的种类有关,所以本文的空间转录组数据 Spots 解卷积结果与真实染色情况吻合,验证了 STDN 算法的准确性。

另外,STDN 模型 Sigmoid 激活函数的输出转化为行和为 1 的概率矩阵(Spots × 细胞类型),矩阵值被解释为将 Spots 分配给对应细胞类型的概率,这是一个 Spot 中细胞属于该类型的比例近似值。将每个 Spot 的细胞类型组成及其比例绘制成饼图投影在对应的组织空间位置上得到图 5 所示饼状散点图。由图 5 可以看出某些区域的某种细胞类型占比较大,具有明显的区域特征,如在相应位置,绿色对应的上皮细胞占比较大,这与图 4(a)的组织学染色图像高度一致。STDN 预测的结果说明了小鼠海马体组织的异质性,更直观说明了细胞类型分布与空间位置有关的观点[4]。

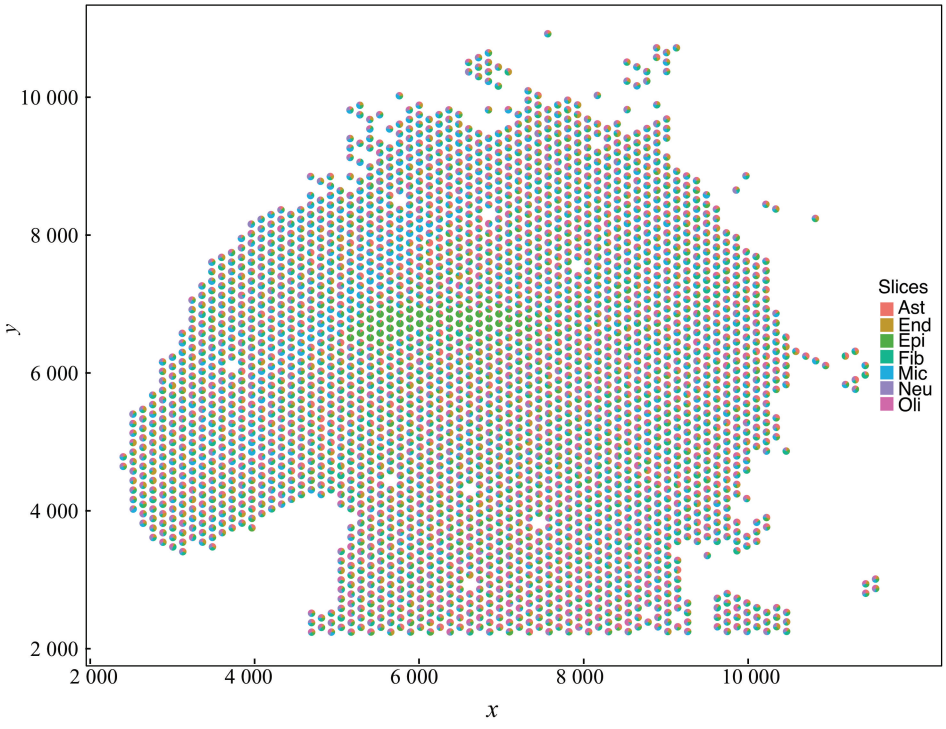

图 5　出生后 7 d 小鼠大脑海马体的 ST 数据解卷积饼状散点图

Fig.5　ST data deconvolution pie scatter plot of mouse brain hippocampus 7 days after birth

注:图中 x,y 均代表 Spot 在组织切片中的空间位置坐标.(扫本文首页二维码见彩图).

2.3　人类胰腺导管腺癌数据异质性分析

2.3.1　人类胰腺导管腺癌数据集

本文分析的人类胰腺导管腺癌数据来自未经治疗的胰腺导管腺癌患者胰腺导管切片,整合 scRNA-seq 数据和 ST 数据进行空间转录组解卷积旨在识别癌变和非癌变区域。其中 scRNA-seq 数据和配对的 ST 数据来自文献[20]和 GEO 数据库,GEO 数据序列号为 GSM3036909, GSM3036911。

2.3.2　实验结果分析

人类胰腺导管腺癌数据实证分析以人类胰腺导管腺癌空间转录组测序的组织切片图像和组织学家基于 H&E 染色(Hematoxylin and eosin (H&E) stains,苏木精和伊红染色)标记的多个组织区域(癌变区、胰腺区、导管区和间质区)[20]为基本事实进行对照,见图 6(a)。将交集处理后的基因表达矩阵及 scRNA-seq 细胞类型注释作为输入训练 STDN 模型,用训练好的模型对人类胰腺导管腺癌 ST 数据进行解卷积,将 Sigmoid 激活函数的输出转化为行和为 1 的概率矩阵,表示 ST 数据每个 Spot 的细胞类型及其比例,绘制饼状散点图,如图 6(b)所示。

由图 6(b)可以观察到切片右上角区域的 Cancer_clone_A 及 Cancer_clone_B 细胞类型占比较大,具有明显的区域特征,这与图 6(a)的癌变区标注位置保持一致;左侧边缘区域导管组织细胞占比较大,这与图 6(a)的导管区标注位置也是一致的。数据分析结果表明人类胰腺导管腺癌 ST 数据解卷积结果与人工标注情况吻合,进一步验证了 STDN 模型解卷积结果的准确性。图 6(b)显示了不同类型在癌变区和正常组织区域呈现出明显不同的空间分布模式,说明了胰腺导管腺癌具有高度的异质性。

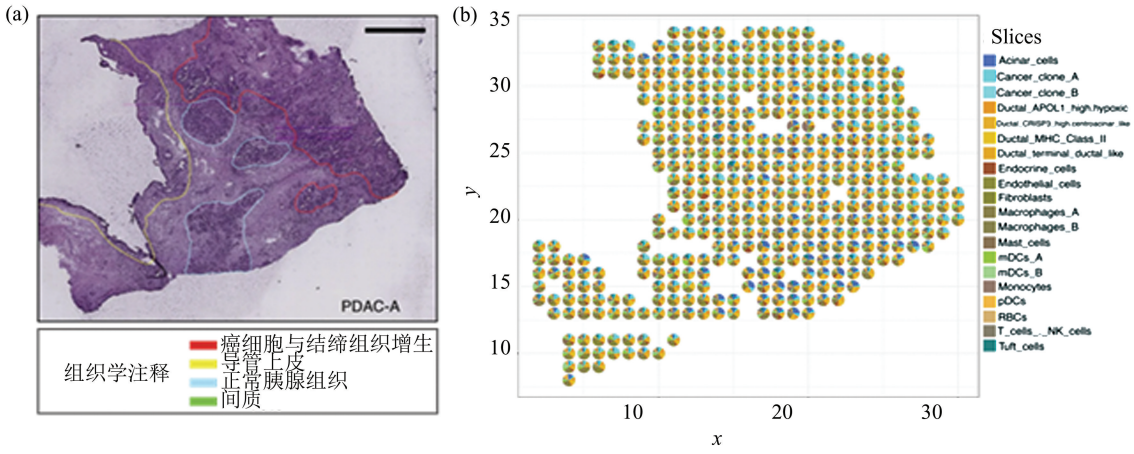

图 6 人类胰腺导管腺癌组织切片及 ST 数据解卷积结果

Fig.6 Human pancreatic ductal adenocarcinoma tissue slice and ST data deconvolution result

注:图中 x,y 均代表 Spot 在组织切片中的空间位置坐标;(a) 组织切片染色及区域标注[9];(b) 解卷积饼状散点图.(扫本文首页二维码见彩图).

正确地识别癌变区至关重要,同时为了更直观地比较解卷积结果与实际切片染色情况,本文将其关联得分投影到 ST 数据组织切片的空间位置,得到如图 7 所示的 20 种细胞类型中与癌变及区域识别相关的 2 种细胞类型的预测结果。与人类胰腺导管腺癌 ST 数据解卷积饼状散点图 6(b)类似,图 7 再次验证了算法解卷积结果的准确性。显然,STDN 模型为组织切片右上角区域赋予了更高的癌变细胞(Cancer_clone_A, Cancer_clone_B)关联得分,意味着这片区域更可能是癌变区域,这与图 6(a)人类胰腺导管腺癌切片 H&E 染色图像及癌变区域标注高度一致。

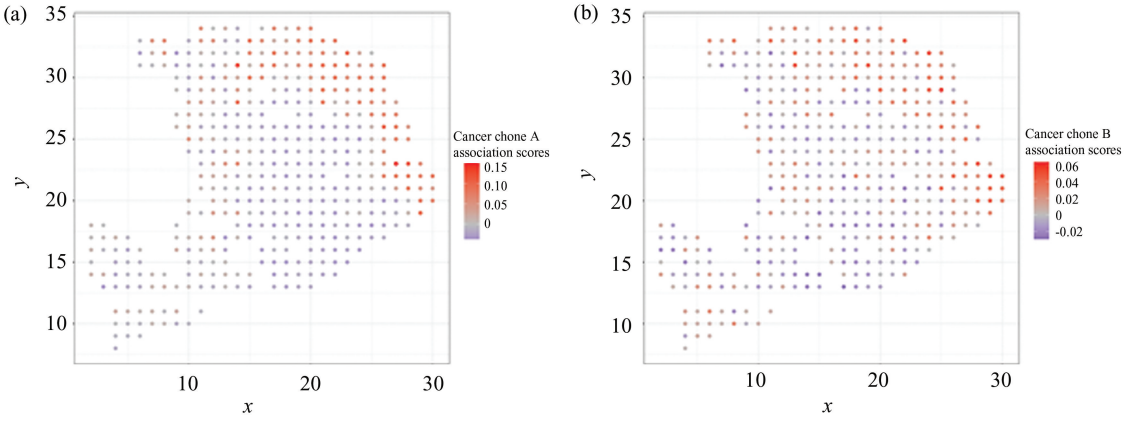

图 7 人类胰腺导管腺癌 ST 数据部分细胞类型关联得分图

Fig.7 Correlation score map of partial cell types in human pancreatic ductal adenocarcinoma ST data

注:图中 x,y 均代表 Spot 在组织切片中的空间位置坐标;(a) 癌症克隆细胞 A 关联得分;(b) 癌症克隆细胞 B 关联得分.(扫本文首页二维码见彩图).

3 结 论

本文充分利用了单细胞转录组数据和空间转录组数据特征的优势,将单细胞转录组数据作为参考信息,引入了卷积神经网络和领域自适应理论,提出了针对空间转录组数据的细胞类型解卷积算法。然后,利用 scRNA-seq 数据及模拟的配套 ST 数据将本算法与其他经典的解卷积算法比较,并利用小鼠大脑海马体数据和人类胰腺导管癌数据进行了实证分析,结果表明,本算法在模拟数据和真实数据上均具有较好的性能。

本文基于 DenseNet 网络和领域自适应理论构建了 STDN 解卷积算法。鉴于不对称转换更灵活,通常能在域自适应任务中产生更好的性能;同时考虑基因测序数据稀疏、高维的特点,STDN 使用非对称转换 CORAL 度量 scRNA-seq 数据和 ST 数据的分布差异,将带有细胞类型标签的 scRNA-seq 数据作为参考,同时最小化细胞类型分类错误、scRNA-seq 数据和 ST 数据的潜在差异,推测 ST 数据中每个捕获位点的细胞

类型组成及比例,达到 ST 数据解卷积的目的。

然而,STDN 具有局限性,也是目前基于 scRNA-seq 参考数据集的解卷积算法存在的共性问题:ST 数据的解卷积过于依赖 scRNA-seq 数据。首先,ST 数据解卷积无法得出 scRNA-seq 数据中没有的细胞类型。其次,scRNA-seq 数据中不准确的细胞类型注释会极大程度影响 ST 解卷积的效果。这是我们未来研究需要解决的问题。另外,理论上,STDN 也可将 ST 数据的空间位置信息迁移到 scRNA-seq 数据的单个细胞上,从而对单细胞数据进行空间重建。无论是 ST 数据解卷积,还是 scRNA-seq 数据空间重建,均是二者信息互相迁移的过程。因此,我们将改进算法,同时达到 ST 数据解卷积和 scRNA-seq 数据空间重建的目的。

致谢:陈子睿、杨博然和何田韵对这项工作的贡献相等。

参考文献(References)

[1]ZHUANG Xiaowei. Spatially resolved single-cell genomics and transcriptomics by imaging[J]. Nature Methods, 2021, 18(1): 18-22. DOI: 10.1038/s41592-020-01037-8.

[2]LUDVIG L, JONAS F, JOAKIM L. Spatially resolved transcriptomics adds a new dimension to genomics[J]. Nature Methods, 2021, 18(1): 15-18. DOI: 10.1038/s41592-020-01038-7.

[3]TANG Xiaoning, HUANG Yongmei, LEI Jinli, et al. The single-cell sequencing: New developments and medical applications[J]. Cell Bioscience, 2019, 9: 53. DOI: 10.1186/s13578-019-0314-y.

[4]LI Yiyun, STANOJEVIC S, GARMIRE L X. Emerging artificial intelligence applications in spatial transcriptomics analysis[J]. Computational and Structural Biotechnology Journal, 2022, 20: 2895-2908. DOI: 10.1016/j.csbj.2022.05.056.

[5]ELOSUA-BAYES M, NIETO P, MEREU E, et al. SPOTlight: Seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomics[J]. Nucleic Acids Research, 2021, 49(9): e50. DOI: 10.1093/nar/gkab043.

[6]DONG Rui, YUAN Guocheng. SpatialDWLS: Accurate deconvolution of spatial transcriptomic data[J]. Genome Biology, 2021, 22: 145. DOI: 10.1186/s13059-021-02362-7.

[7]CABLE D M, MURRAY E, ZOU L S, et al. Robust decomposition of cell type mixtures in spatial transcriptomics[J]. Nature Biotechnology, 2022, 40(4): 517-526. DOI: 10.1038/s41587-021-00830-w.

[8]KLESHCHEVNIKOV V, SHMATKO A, DANN E, et al. Cell2location maps fine-grained cell types in spatial transcriptomics[J]. Nature Biotechnology, 2022, 40(5): 661-671. DOI: 10.1038/s41587-021-01139-4.

[9]MA Ying, ZHOU Xiang. Spatially informed cell-type deconvolution for spatial transcriptomics[J]. Nature Biotechnology, 2022, 40: 1349-1359. DOI: 10.1038/s41587-022-01273-7.

[10]ANDERSSON A, BERGENSTRAHLE J, ASP M, et al. Single-cell and spatial transcriptomics enables probabilistic inference of cell type typography[J]. Communications Biology, 2020, 3: 565. DOI: 10.1038/s42003-020-01247-y.

[11]SONG Qianqian, SU Jing. DSTG: Deconvoluting spatial transcriptomics data through graph-based artificial intelligence[J]. Briefings in Bioinformatics, 2021, 22(5): bbaa414. DOI: 10.1093/bib/bbaa414.

[12]SUN Dongqing, LIU Zhaoyang, LI Taiwen, et al. STRIDE: Accurately decomposing and integrating spatial transcriptomics using single-cell RNA sequencing[J]. Nucleic Acids Research, 2022, 50(7): e42. DOI: 10.1093/nar/gkac150.

[13]LOPEZ R, LI Baoguo, KEREN-SHAUL H, et al. DestVI identifies continuums of cell types in spatial transcriptomics data[J]. Nature Biotechnology, 2022, 40: 1360-1369. DOI: 10.1038/s41587-022-01272-8.

[14]张柳. 基于子空间对齐的单细胞转录组测序数据的细胞类型预测研究[D]. 武汉: 华中师范大学, 2021. DOI: 10.27159/d.cnki.ghzsu.2021.001708.
ZHANG Liu. Cell type prediction of single-cell transcriptome sequencing data based on subspace alignment[D]. Wuhan: Central China Normal University, 2021. DOI: 10.27159/d.cnki.ghzsu.2021.001708.

[15]BAE S, NA K J, KOH J, et al. CellDART: Cell type inference by domain adaptation of single-cell and spatial transcriptomic data[J]. Nucleic Acids Research, 2022, 50(10): e57. DOI: 10.1093/nar/gkac084.

[16]SUN Baochen, SAENKO K. Deep coral: Correlation alignment for deep domain adaptation[C]//Computer Vision-ECCV 2016 Workshops. Amsterdam, The Netherlands, Springer International, 2016:443-450. DOI:10.1007/978-3-319-49409-8_35.

[17]FUGLEDE B, TOPSOE F. Jensen-shannon divergence and hilbert space embedding[C].// International Symposium on Information Theory, ISIT 2004. Chicago, IL, USA: Proceedings, 2004, 31. DOI: 10.1109/ISIT.2004.1365067.

[18]LI Bin, ZHANG Wei, GUO Chuang, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution[J]. Nature Methods, 2022, 19(6): 662-670. DOI: 10.1038/s41592-022-01480-9.

[19]JOGLEKAR A, PRJIBELSKI A, MAHFOUZ A, et al. A spatially resolved brain region-and cell type-specific isoform atlas of the postnatal mouse brain[J]. Nature Communications, 2021, 12(1): 463. DOI: 10.1038/s41467-020-20343-5.

[20]MONCADA R, BARKLEY D, WAGNER F, et al. Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas[J]. Nature Biotechnology, 2020, 38(3): 333-342. DOI: 10.1038/s41587-019-0392-8.