DOI:10.11918/202509121

一种单细胞转录组与免疫组库的整合方法

何 昭,亓 晶,靳水林

(哈尔滨工业大学 数学学院,哈尔滨 150001)

摘 要:单细胞转录组测序(scRNA-seq)与T细胞受体免疫组库测序(scTCR-seq)是解析免疫细胞特性的两大关键技术,分别从基因表达和抗原识别维度揭示免疫系统的复杂性。然而,传统分析方法多局限于单一模态,难以有效整合两个组学所提供的互补信息。为突破这一局限,实现跨组学数据的高效融合,提出一种新型的数据整合架构scRTIA(single cell RNA and TCR integrative analysis)。该模型基于深度学习理论,以多模态变分自编码器与Transformer为核心架构,将scRNA-seq基因表达矩阵与scTCR-seq的TCR序列特征协同嵌入统一的低维潜在空间,从而构建出能同时保留转录组特征与免疫组库信息的融合细胞表征。在真实数据集上的实验验证表明,scRTIA所构建的细胞表征在细胞亚群识别方面表现出显著优越的分辨能力,能够发现传统方法难以识别的、具有特定功能状态的稀有T细胞群体。本研究通过有效深度融合转录组与免疫组库信息,突破了单模态分析的瓶颈,实现了对T细胞身份和功能的多维度刻画,在免疫相关疾病研究和精准医疗领域具有应用价值。

关键词:单细胞转录组;单细胞免疫组库;数据整合;Transformer;变分自编码器

中图分类号:O213.9 **文献标志码**:A **文章编号**:0367-6234(2025)12-0294-10

An integrated approach for single-cell transcriptome and immune repertoire

HE Zhao, QI Jing, JIN Shuilin

(School of Mathematics, Harbin Institute of Technology, Harbin 150001, China)

Abstract: Single-cell RNA sequencing (scRNA-seq) and single-cell T-cell receptor sequencing (scTCR-seq) are pivotal for deciphering immune cell characteristics, offering complementary insights into the immune system's complexity through gene expression and antigen recognition, respectively. However, conventional analytical methods are often confined to a single modality, hindering the effective integration of complementary information from the two omics. To overcome this limitation and achieve efficient integration of cross-omics data, this study proposes a novel data integration framework named scRTIA (single-cell RNA and TCR integrative analysis). Based on deep learning theory, the model employs a multimodal variational autoencoder and a Transformer as its core architecture, jointly embedding the scRNA-seq gene expression matrix and scTCR-seq TCR sequence features into a unified low-dimensional latent space, thereby constructing a fused cell representation that simultaneously preserves transcriptomic features and immune repertoire information. Experimental validation on real datasets demonstrates that the cell representations generated by scRTIA exhibit significantly superior resolution in identifying cell subpopulations, enabling the discovery of rare T-cell populations with specific functional states that are difficult to detect using traditional methods. By effectively integrating transcriptomic and immunome data, our work transcends the limitations of single-modal analysis and enables a multidimensional characterization of T-cell identity and function, offering valuable insights for immune-related diseases research and precision medicine.

Keywords: single-cell transcriptome; single-cell immune repertoire; data integration; Transformer; variational antoencoder

单细胞组学兴起于2009年提出的单细胞转录组测序技术(single-cell RNA sequencing, scRNA-seq),改变了之前批量转录组测序技术只能对整个批次细胞进行低分辨率转录组测序的状况,实现了对单个细胞的全转录组测序。但是,早期的单细胞转录组技术通量低,每次只能测序几十到几百个细胞。经过近几年的发展,在2014年逐渐出现了高通量单细胞转录组测序技术,实现了单细胞测序技术的降本增效。时至今日,单细胞转录组测序的通量已经可以达到数万甚至数十万个细胞[1]。

收稿日期:2025-09-30;**录用日期**:2025-11-10;**网络首发日期**:2025-11-19
网络首发地址:https://link.cnki.net/urlid/23.1235.T.20251119.1521.006
基金项目:国家自然科学基金(12301623,62271173,62531006);黑龙江省自然科学基金(LH2024A003,JQ2023A003);黑龙江省博士后科研启动金(LBH-Z23020);中国博士后科学基金(GZC20233473);中央高校基本科研业务费专项资金(HIT.DZJJ.2024043)
作者简介:何 昭(2003—),男,博士研究生;靳水林(1980—),男,博士生导师
通信作者:亓 晶,qijing@hit.edu.cn;靳水林,jinsl@hit.edu.cn

随着单细胞转录组的出现，越来越多的其他单细胞组学技术也开始蓬勃发展，如单细胞染色质可及性，关注单细胞蛋白质表达情况的数据等，均为研究人员从不同视角研究单细胞组学提供了技术支持[2]。在这些多组学技术中，T 细胞免疫组库测序技术（single-cell T cell receptor sequencing，scTCR-seq）是一项重要的技术。这项技术主要关注生物体中一种重要的免疫细胞——T 细胞上的免疫受体（TCR）信息，因为这部分信息直接体现了 T 细胞的特异性[3]。相比 scRNA-seq 与染色质可及性数据，蛋白组数据的整合研究，scRNA-seq 与 scTCR-seq 的整合相关研究较少，目前为止，大致可以分为基于统计学的方法与基于深度学习的方法，这些方法各有特点且均有一些不足之处。

在基于统计学的方法中，CoNGA 模型[4]使用克隆型邻域图分析，识别转录组文件与 TCR 之间的相关性，并且通过定义“CoNGA 评分”进行量化；Tessa 模型[5]使用参数贝叶斯层次模型，整合 TCR 数据和 T 细胞的基因表达数据，评估 TCR 序列对 T 细胞表型的影响。这两篇早期文章奠定了 scTCR-seq 与 scRNA-seq 整合研究的基础，阐述了该研究方向的重要性与可行性，但是，这两篇文章均假设具有相似基因表达情况的 T 细胞具有相似的 TCR 序列信息，而 T 细胞分化特异性复杂，此生物学假设不完全准确。自此，通过某种方法实现弱生物学假设下的双模态整合成为了研究重点。随着深度学习技术的发展，逐渐有研究转向使用神经网络对这两个模态数据进行整合。scNAT 模型[6]基于联合卷积神经网络的变分自编码器架构对双模态数据进行整合，将双模态数据映射入一个联合潜在空间；mvTCR 模型[7]使用 Transformer 模型对字符型免疫组库数据进行编码，之后利用混合变分自编码器进行整合。这两种方法需要较少的生物学假设，但是，前者只考虑了双模态整合之后的潜在表示，对 T 细胞的免疫组数据分布考虑不足，而后者更关心 TCR 的 CDR3 序列信息，对编码 v,d,j 基因的重排情况考虑较少。近日提出的 MIST 方法[8]充分考虑了多模态信息，但是，其免疫组数据只精确到了单细胞程度，没有对 TCR 双链进行分析，并且，没有对 T 细胞免疫组库数据分布高度离散的特性进行处理。

因此，本研究提出了 scRTIA（single cell RNA and TCR integrative analysis）模型，利用 Transformer[9]，基于高斯混合模型的变分自编码器（GMM-VAE）等神经网络架构，结合概率生成模型进行整合。不同于之前方法将免疫组看作一个整体模态的思路，本方法更多地从生物学角度出发，利用 TCR 由双链构成这个生物学事实，实现从免疫双链开始的对 scRNA-seq 和 scTCR-seq 数据的整合分析，从而使得最终的细胞潜在表示可以对 T 细胞双链的差异进行区分，也可以对 T 细胞群体进行更精细的刻画，使得方法具有发现新生物学现象的潜力。并且，本研究使用 T 细胞转录组和免疫组的真实配对数据集，验证方法的可行性。

1　单细胞免疫组库与转录组

1.1　T 细胞免疫组库

T 细胞是一类重要的特异性免疫细胞，主要执行免疫系统的适应性免疫反应，可以识别攻击被感染的细胞或者癌细胞，并且调控其他免疫细胞参与免疫活动。TCR 是镶嵌在 T 细胞表面的一个蛋白质复合物，可以特异性识别抗原，从而使得 T 细胞识别目标，被激活以及发挥其免疫作用。

TCR 一般由双链组成，大部分 T 细胞的双链为 α 链和 β 链，这类 T 细胞被称为 $\alpha\beta$ T 细胞，少部分 T 细胞是 γ 链和 δ 链组成的，这类 T 细胞被称为 $\gamma\delta$ T 细胞。对于这些链来说，都是多基因编码的，并且均分为 V 区（可变区）和 C 区（恒定区），3 类不同的基因片段。一般地，v,d,j 基因，编码 β 链和 δ 链的 V 区，而 v 和 j 编码 α 链和 γ 链的 V 区。T 细胞成熟过程中，在胸腺中进行 V 区基因的重排，从而导致了 T 细胞的高度多样性[10]。

在 TCR 上，CDR3 区（互补决定区 3）是最核心和多变的部分，直接决定了免疫细胞可以识别哪一种特定的抗原。CDR3 区位于 V 区，由 v,d,j 基因的重排进行编码，并且伴随核苷酸的随机插入或缺失，这就导致不同 T 细胞 CDR3 区的氨基酸序列几乎都是不同的。由此，T 细胞免疫组库的主要内容就有每一个细胞的 TCR 双链，具体信息包含两部分，一部分是每一条链上的 CDR3 区氨基酸序列，另一部分是编码这些氨基酸序列对应使用的 v,d,j 基因的类别。其存储使用通用的国际 AIRR 社区指定的标准化格式进行[11]，如 10x 等商业化平台均采用这种存储格式进行数据整理。T 细胞免疫组库均为字符型序列以及类别型基因信息，而非传统的欧几里得型数据，这也给其处理带来困难。如何对这种数据进行统计建模，或者使用神经网络进行处理，是此类数据研究的重点内容。

1.2　单细胞转录组

由中心法则，遗传信息由 DNA 转录到 RNA，再由 RNA 翻译成蛋白质去执行细胞的生物学功能。细胞内部的 mRNA 既代表了细胞内的基因表达信息，又可以用于推测细胞的蛋白质表达，因此，单细

胞转录组处于单细胞组学的核心地位。目前,标准的单细胞转录组数据格式为基因表达矩阵,即行代表细胞,列代表基因,而矩阵中的每一项代表某个细胞某个基因的表达量。

2 方法构建

本方法从 TCR 的结构出发,按照 α 链和 β 链(γ 链和 δ 链)组成 TCR 免疫组完整信息,免疫组数据与转录组数据整合得到最终潜在表示的流程,模块化地进行整合,最终可以得到 α 单链(γ 单链)的分布、β 单链(δ 单链)的分布、免疫组信息分布以及免疫组与转录组的整合数据分布。整个方法流程见图 1,以下按照模块介绍方法各部分的细节。

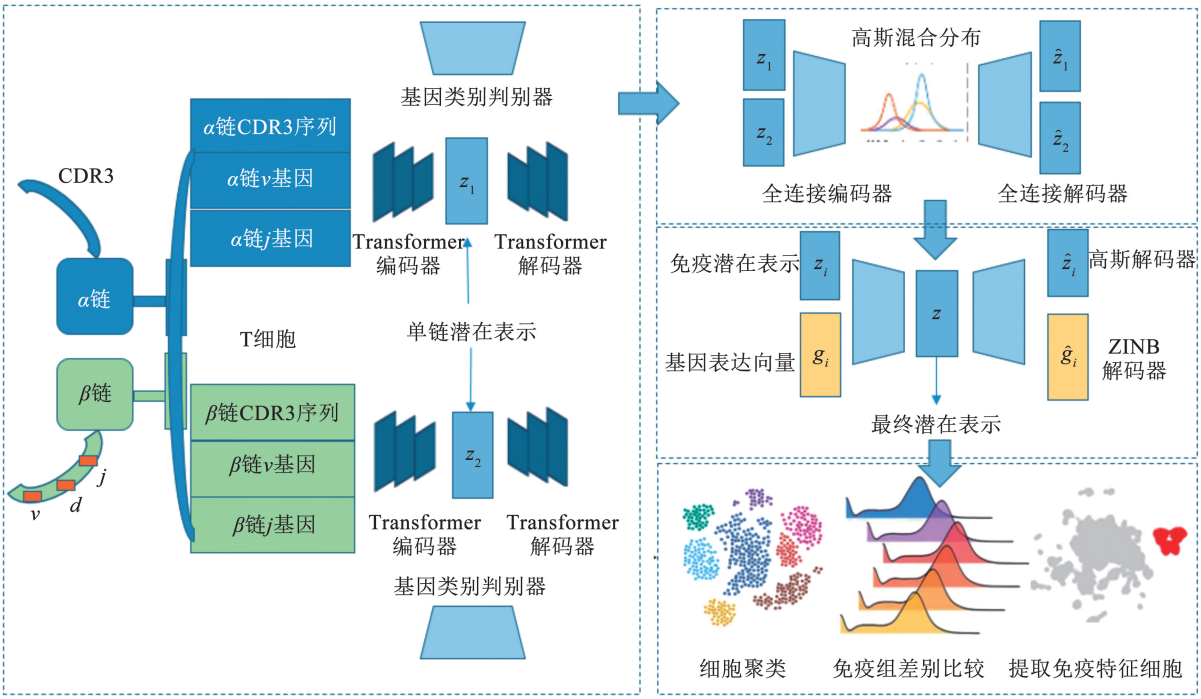


图 1 方法流程

Fig. 1 Methodology flowchart

2.1 基于 Transformer 的单链嵌入模块

Transformer 是处理序列型数据的重要神经网络架构,不同于循环神经网络(RNN),其对前后序列的关联性主要基于自注意力机制。本研究使用 Transformer 架构搭建编码器,对字符型氨基酸序列进行连续数值嵌入。具体地,首先,建立氨基酸字母与整数的词汇表映射,使用 0 将序列扩充到最大序列长度。之后,通过可学习嵌入层,获取氨基酸序列的对应数值语义向量。为了将氨基酸序列的顺序信息注入语义向量,使用正弦余弦编码生成位置向量,即

$$\mathrm{PE}_{(\mathrm{pos},2i)} = \sin \frac{\mathrm{pos}}{10\ 000^{2i/d_{\mathrm{model}}}} \tag{1}$$

$$\mathrm{PE}_{(\mathrm{pos},2i+1)} = \cos \frac{\mathrm{pos}}{10\ 000^{2i/d_{\mathrm{model}}}} \tag{2}$$

式中:PE 表示位置编码;pos 代表元素在序列中的位置;i 为维度索引,满足 $0 \leqslant i \leqslant d_{\mathrm{model}}/2$,$d_{\mathrm{model}}$ 为嵌入的维数;10 000 为位置编码常用参数,来自 Transformer 论文中的默认参数。

对于处理过后的向量,将其通过若干个 Transformer 编码层。在每一层 Transformer 编码层中,向量首先利用自注意力机制

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V} \tag{3}$$

得到氨基酸数值向量,其中,$\boldsymbol{Q}$ 为查询矩阵,$\boldsymbol{K}$ 为键矩阵,$\boldsymbol{V}$ 为值矩阵,d_K 为每个键向量的维度,用于缩放注意力,使得注意力计算更加稳定。之后通过前馈神经网络,并且将输出与原始向量通过残差连接得到此层的输出,在通过若干层之后,得到编码的潜在向量。

为了充分利用免疫组库信息,得到的潜在向量分别通过两个解码部分,一部分是两个类别解码器,用于预测潜在向量对应细胞使用的重排 v,j 基因类型,确保潜在表示蕴含基因重排信息;另一部分是 Transformer 的解码器,通过该解码器重构细胞的原始氨基酸序列,确保潜在表示蕴含氨基酸序列信息。其中,类别解码器均使用单层全连接层,输出为预测的 v,j 基因各种类别的嵌入表示概率,对应的损失

函数为

$$L_v = -\sum_{i=1}^{M_1} v_i \log \hat{v}_i \tag{4}$$

式中：v_i 为真实的 v 基因类别信息，$\hat{v}_i$ 为预测的基因类别概率，M_1 为 v 基因类别总数。类似地，

$$L_j = -\sum_{i=1}^{M_2} j_i \log \hat{j}_i \tag{5}$$

式中：j_i 为真实 j 基因的类别信息，$\hat{j}_i$ 为预测的基因类别概率，M_2 为 j 基因类别总数。对于氨基酸序列重构部分，使用交叉熵损失

$$L_{\text{recon}} = \frac{-1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it} \log p_{it} \tag{6}$$

式中：N 为批次大小，T 为序列长度。因此，总损失函数为 $L_{\text{chain}} = \alpha_1 L_{\text{recon}} + \beta_1 L_v + \gamma_1 L_j$，其中，$\alpha_1$，$\beta_1$，$\gamma_1$ 为可调的权重参数，本文初始设置为 $\alpha_1 = \beta_1 = \gamma_1 = 1$。

2.2　基于高斯混合潜在分布的免疫组双链整合变分自编码器

对于基本的变分自编码器架构（variational autoencoder，VAE），一般假定其潜在空间为标准高斯分布，然而，对于一些复杂的数据分布，一个普通的高斯分布难以拟合复杂的数据分布情况。对于免疫组数据而言，由于免疫克隆型的高度特异性，其类别众多，分布离散。因此，采用高斯混合模型（Gaussian Mixture Model，GMM）修正的变分自编码器，对 2.1 节得到的双链表示进行整合，进而得到免疫组库的潜在表示。

以 $\alpha\beta$ T 细胞的处理为例，$\gamma\delta$ T 细胞的处理是完全类似的。设在 2.1 节中得到的双链数值嵌入分别为 z_1，z_2，将二者拼接起来得 $\boldsymbol{z}_{\text{concat}} = (z_1, z_2)$，目的是将其通过 GMM 变分自编码器得到双链的融合潜在表示。由于高斯混合分布具有混合系数、均值向量和协方差矩阵 3 个参数，其的组合决定了数据分布。虽然在网络学习中，这 3 个参数可以通过数据学习得到更新，但是，一个合理的初始化可以帮助参数更好地逼近分布。因此，本研究采用预训练与高斯混合训练通过 EM 算法结合的方法进行网络训练。

具体地，将拼接向量 $\boldsymbol{z}_{\text{concat}}$ 首先通过一个以标准高斯分布为潜在空间的变分自编码器，其损失函数为常见的向量重构损失和标准高斯 KL 散度，经过预训练之后，可以得到一个初始的数据潜在空间分布。在得到潜在表示后，给定聚类中心数目，通过 EM 算法计算每个点属于各类中心的后验概率，之后更新聚类中心，逐步迭代下去得到潜在空间作为高斯混合分布的初始参数 μ_k，Σ_k，π_k。

得到预训练的初始参数之后，利用其作为 GMM-VAE 的训练起点，进行双链整合训练。具体地，将 $\boldsymbol{z}_{\text{concat}} = (z_1, z_2)$ 作为编码器原始输入，映射到高斯混合模型的潜在空间，之后再利用解码器重构输入向量，得到重构向量 $\hat{\boldsymbol{z}}_{\text{concat}} = (\hat{z}_1, \hat{z}_2)$，对于重构项，采用均方误差作为损失函数，即

$$L_1 = \frac{1}{N}\sum_{i=1}^{N} (z_1 - \hat{z}_1)^2 \tag{7}$$

$$L_2 = \frac{1}{N}\sum_{i=1}^{N} (z_2 - \hat{z}_2)^2 \tag{8}$$

其中 N 为批次大小。

对于变分自编码器的 KL 散度项，使用编码分布与高斯混合先验的 KL 散度进行训练。具体地，

$$L_{\text{KLGMM}} = E_{z\sim q(z|x)}(\ln q(z|x) - \ln p(z))$$

其中 $\ln p(z) = \ln\left(\sum_{k=1}^{K} \pi_k N(z;\mu_k,\Sigma_k)\right)$，而 $\ln q(z|x) = \ln N(z;\mu_\phi,\Sigma_\phi)$，由其显示表达有

$$E_q[\ln q(z|x)] = -\frac{1}{2}(d\ln 2\pi + \ln|\Sigma_\phi| + d) \tag{9}$$

而另一项没有闭解，故采用蒙特卡洛估计近似逼近这个积分式，即在大量采样后，用求和式逼近这个积分。其中，d 为空间维数，总损失函数为 $L = \alpha_2 L_{\text{KLGMM}} + \beta_2 L_1 + \gamma_2 L_2$，$\alpha_2$，$\beta_2$，$\gamma_2$ 为可调的参数。一般地，由于大部分 T 细胞的特异性体现在 β 链上，初始设置为 $\alpha_2 = 0.1$，$\beta_2 = 0.2$，$\gamma_2 = 0.7$。

2.3　基于概率生成模型的双模态整合

在本模块中，采用变分自编码器的思路，通过编码器将数据映射到潜在空间，之后通过重构进行训练。这里设 $q_\phi(z|x)$ 是近似后验分布，$p_\theta(x|z)$ 是解码器重构，为了考虑双模态，设基因表达信息为 X_{gex}，免疫组信息为 X_{imm}，潜在表示为 Z，则全概率模型为

$$P(X_{\text{gex}}, X_{\text{imm}}, Z) = P(X_{\text{gex}}|Z)P(X_{\text{imm}}|Z)P(Z) \tag{10}$$

此模块最终的优化目标函数为

$$L_{\text{final}} = E_{q(z|x)}[\log p(x_{\text{gex}}|z) + \log p(x_{\text{imm}}|z)] - D_{\text{KL}}(q(z|x)||p(z)) \tag{11}$$

其中 $x = (x_{\text{gex}}, x_{\text{imm}})$。

对于基因表达信息，为了拟合基因表达的分布以及考虑单细胞转录组数据具有技术零值的特点，使用零膨胀负二项分布（ZINB）进行建模生成，ZINB 分布的概率密度函数为

$$P_{\text{ZINB}} = \pi \cdot \delta_0(x) + (1 - \pi_{ij}) \cdot \text{NB}(x|\mu,\phi) \tag{12}$$

式中：π 为零膨胀概率，NB 为负二项分布，μ 为负二项分布均值，ϕ 为离散度参数，这些参数通过解码器得到，则 $L_{\text{ZINB}} = -\log P_{\text{ZINB}}(x|\mu,\phi)$。

对于免疫组数据，采用高斯分布的连续解码，对应的损失函数为

$$L_{\text{imm}} = \frac{1}{N}\sum_{i=1}^{N} \| x_{\text{imm}}^{i} - \text{decoder}(z^{i}) \|^{2} \quad (13)$$

KL 散度项为 $D_{\text{KL}}(q(z \mid x) \,||\, N(0,I)) = \frac{1}{2}\sum_{j=1}^{J}(\sigma_j^2 + \mu_j^2 - 1 - \log \sigma_j^2)$。最终的此模块损失函数为

$$L_{\text{final}} = \omega_{\text{gex}} \cdot L_{\text{ZINB}} + \omega_{\text{imm}} \cdot L_{\text{imm}} + \omega_{\text{KL}} \cdot D_{\text{KL}} \quad (14)$$

其中 $\omega = (\omega_{\text{gex}}, \omega_{\text{imm}}, \omega_{\text{KL}})$ 是可调的权重参数，本研究着重体现免疫组库数据对转录组数据分布的扰动效果，故初始设置为 $\omega = (1, 0.2, 0.01)$。

3 实验及结果分析

3.1 数据说明

配对的人类 T 细胞样本可以通过网站直接获取（https://www.10xgenomics.com/），其中的免疫组数据集均按照 AIRR 标准格式进行存储，并且可以通过开源 python 模块 scirpy 直接进行读取。不同的生物条件下小鼠的单细胞转录组测序与免疫组库测序技术来源于 NCBI 数据库 GSE275982 号的公开测序数据集。本研究使用方法代码公开于 https://github.com/Jinsl-lab/scRTIA。

由于测序技术限制，得到的原始基因表达数据与免疫组库数据均受到无效数据与噪声等因素的干扰。在使用数据进行算法验证之前，需要首先对原始数据进行预处理。针对免疫组数据而言，其测序数据为基因类别与 CDR3 区序列，其中的每个信息均在本研究的整合方法中起到重要作用，因此，本研究只保留那些具有完整 CDR3 区序列、α 链 v,j 基因类别 β 链 v,d,j 基因类别的单链，将其他有缺失信息的单链删除。除此之外，只保留全长（fulllength = TRUE），可生成单链（productive = TRUE）的链。在将 TCR 拼成一个细胞时，有时一个细胞只能找到一个单链，而有的细胞可能会在某些链上出现冗余，这两种情况对训练均有影响，并且这些细胞在数据集中均有一定比例的分布，本研究删除掉这些链型，只保留可以正好组成一个单细胞的匹配双链，及只保留占数据集中多数的单链配对 T 细胞，之后通过神经网络进行方法训练。

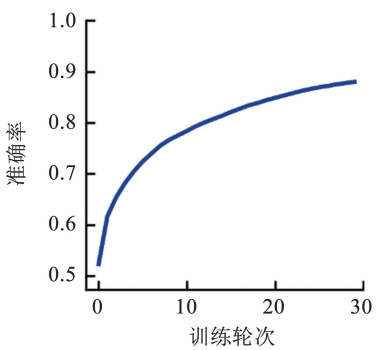
针对单细胞转录组数据，其格式可以是转录组测序的.h5 文件或者.mtx 格式的基因表达矩阵。在预处理部分，首先，通过控制单细胞中的线粒体比例，删去线粒体含量过高、极有可能衰老凋亡的细胞，之后删除在小于 3 个细胞中表达的基因，并且去除掉表达基因数小于 200 的细胞，因为其可能是测序得到的空液滴或者死亡细胞。

3.2 实验结果

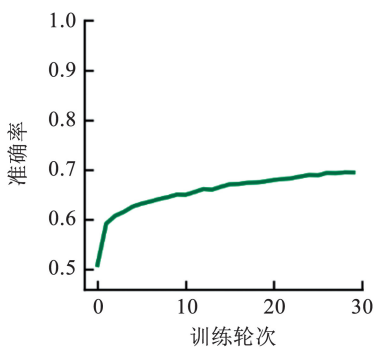
3.2.1 scRTIA 可以充分提取免疫组库信息

由于本方法首先对免疫组库信息进行了提取与整合，对于免疫组库信息的数值嵌入应该整体包含大部分免疫组库信息，否则在方法的下游处理时，准确率会被大幅影响。

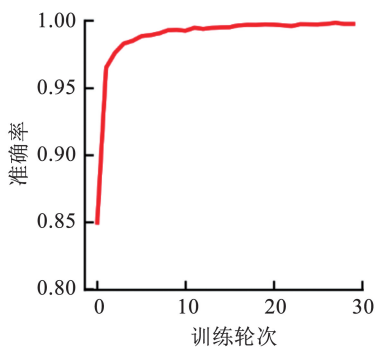
使用来自 4 个不同捐献者的 CD8 + T 细胞的 10x 数据集进行测试，如图 2（a）~（c）所示，随着训练轮次的增加直到神经网络触发早停，方法在免疫组数据的 CDR3 序列重构，v 基因 j 基因类别预测的准确率方面，均随训练轮次的增加而增加，说明通过 Transformer 架构得到的潜在表示在一定程度上保留了主要的免疫组库信息。以 β 链为例，由图 2（d）、（e）可知，含量较多的 v,j 基因在其潜在空间上的分布。在潜在空间上，由于 v 基因的高变性，其分布除了某些基因型较为聚集之外，其他的基因型相对混合分布，而 j 基因相对保守。因此，在潜在空间上，按照不同基因型相对更聚集。



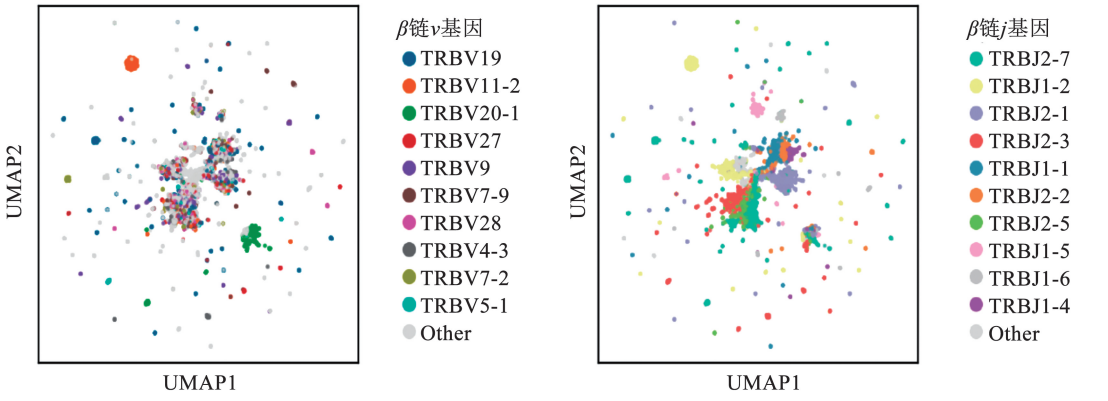
(a) 序列重构准确率



(b) v基因预测准确率



(c) j基因预测准确率



(d) β链上出现频率前十位的v基因分布　　(e) β链上出现频率前十位的j基因分布

图 2　scRTIA 可以提取免疫组库信息

Fig. 2　scRTIA extraction of immune-repertoire features

此外，本研究发现，通过添加 v，j 基因的类别判别损失函数，在训练结果大体相似的情况下，可以显著地减少神经网络的训练时间。如表 1 所示，在不添加 v，j 基因的判别损失时，训练时间均在 11 ~ 13 min，而在添加了 v，j 基因的判别损失之后，训练时间均显著缩短至 2 ~ 6 min。对此，在生物学上可以进行一些解释，即 CDR3 区的氨基酸由 v，d，j 基因编码产生，CDR3 区序列由密码子表翻译得到，所不同的只是随机插入和缺失，由此，v，j 基因的信息在一定程度上均体现了 CDR3 区的序列信息，故加速了神经网络的训练速度。

表 1　是否添加判别项的训练时间比较

Tab. 1　Comparison of training time with and without discriminative item

捐赠者	添加判别项时间/s	不添加判别项时间/s
1	208	830
2	404	816
3	183	685
4	124	785

3.2.2　scRTIA 可以在潜在空间中保持克隆型的类别聚集

在 T 细胞中，克隆型指具有相同或者相似 CDR3 氨基酸序列的 T 细胞，或者具有相同 v，d，j 重排方法的 T 细胞群体。在生物学中，分析 T 细胞的克隆型对分析免疫应答[12]、针对不同疾病的免疫特点[13]、设计肿瘤免疫治疗[14]、分析免疫变化[15]等重要的生物学问题起到关键作用。因此，在整合后的潜在表示中，保持 T 细胞的克隆型分布，对生物学分析十分重要。

事实上，由于 T 细胞的免疫组库数据，与其对应的转录组数据并不是强相关的组学数据，即通过其中某一个组学数据去较为准确地推断另一个组学数据的分布情况是基本不可能的，所以，如图 3(a) 所示，考虑了含量最多的前十位克隆型的分布情况，在转录组数据的潜在分布上，不同的克隆型虽然大体上体现出某种聚集效果，但是，总的来看还是难以体现免疫组的特异性。

在将数据集中 T 细胞的免疫组库信息进行提取整合，得到数据集在免疫组库下的潜在表示空间之后，由图 3(b) 可以看出，相同克隆型的细胞在此潜在空间上出现了部分聚集效应，由于 T 细胞的分布高度离散，其分布大多聚成小但密集的类别。

再按照方法流程，将 T 细胞的免疫组库信息与转录组信息结合，在之后的整合潜在空间中，由图 3(c) 可以看出，相同的 T 细胞克隆型基本被分配到一个聚集的类别中，而不是像免疫组库数据分布那样离散。事实上，T 细胞的分布受到克隆型种类的影响，在潜在空间中比较离散，而转录组数据分布通常比较连续。因此，二者整合后的数据分布会体现出两个模态的共同特点，即有的部分呈现连续的分布状态，而有的离散小类别在潜在空间各自独立分布。

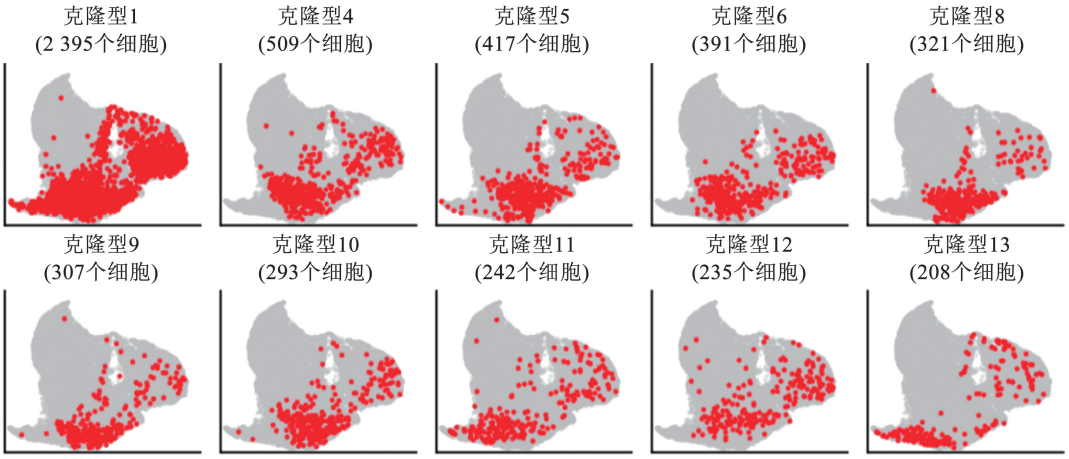
3.2.3　scRTIA 可以部分利用免疫特性划分新细胞类别

对于单细胞转录组，其细胞聚类过程完全依靠于每一个细胞的基因表达信息，这种分类虽然可以体现细胞的表达特性，但对于其他模态的信息利用不足。通过整合免疫组信息，可以在体现转录组特性的同时，依据其免疫组特性进行分类，进而综合研究细胞的特性。

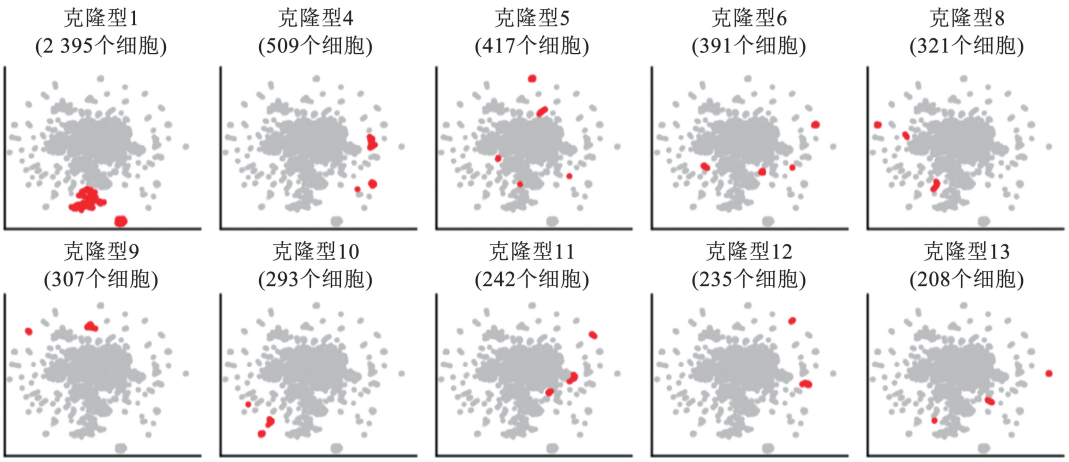
由图 4(a) 可以看出，在潜在空间聚类之后聚出了不同类别，这些类别在原始转录组空间中虽有某些对应的分类趋势，但是无法观察到明显的分类。

其中,不同类别细胞的类型混杂现象比较严重,这是因为转录组数据几乎不能体现过多的免疫组信息,或者说这两个模态是弱关联模态。

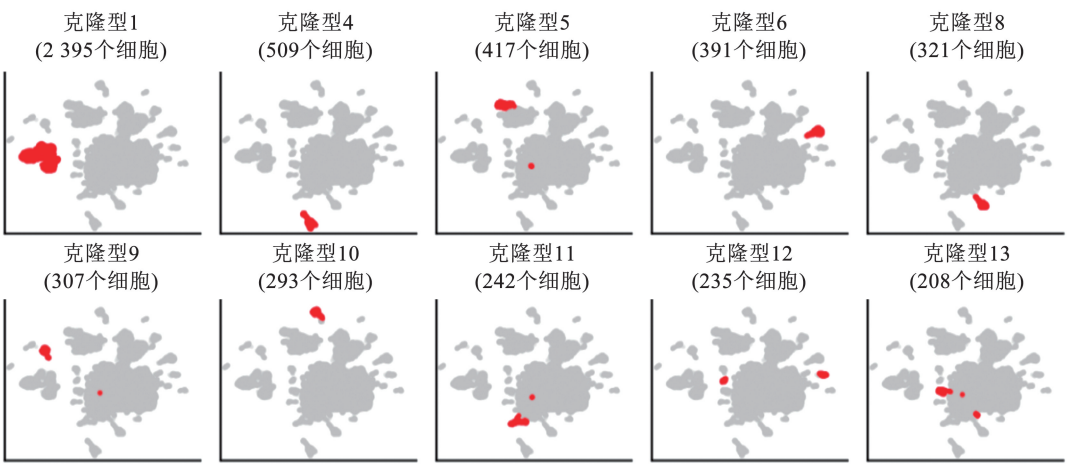
为了探究不同类别的免疫特性,选择了 CDR3 区氨基酸链长度这个特性。对于 TCR,CDR3 区氨基酸链的长度是一个重要的指标,而通过山脊图对潜在空间划分的不同类别细胞对应的免疫组库 CDR3 链的长度进行分布可视化,如图 4(b)所示,不同的类别中,其 CDR3 序列长度分布差别明显。



(a) 含量最多的前十位克隆型在转录空间上的分布



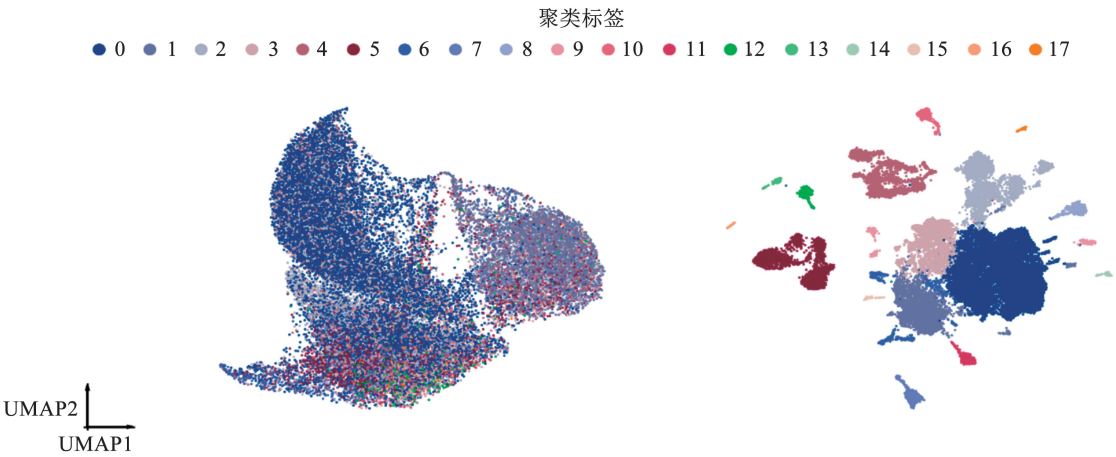
(b) 含量最多的前十位克隆型在免疫空间上的分布



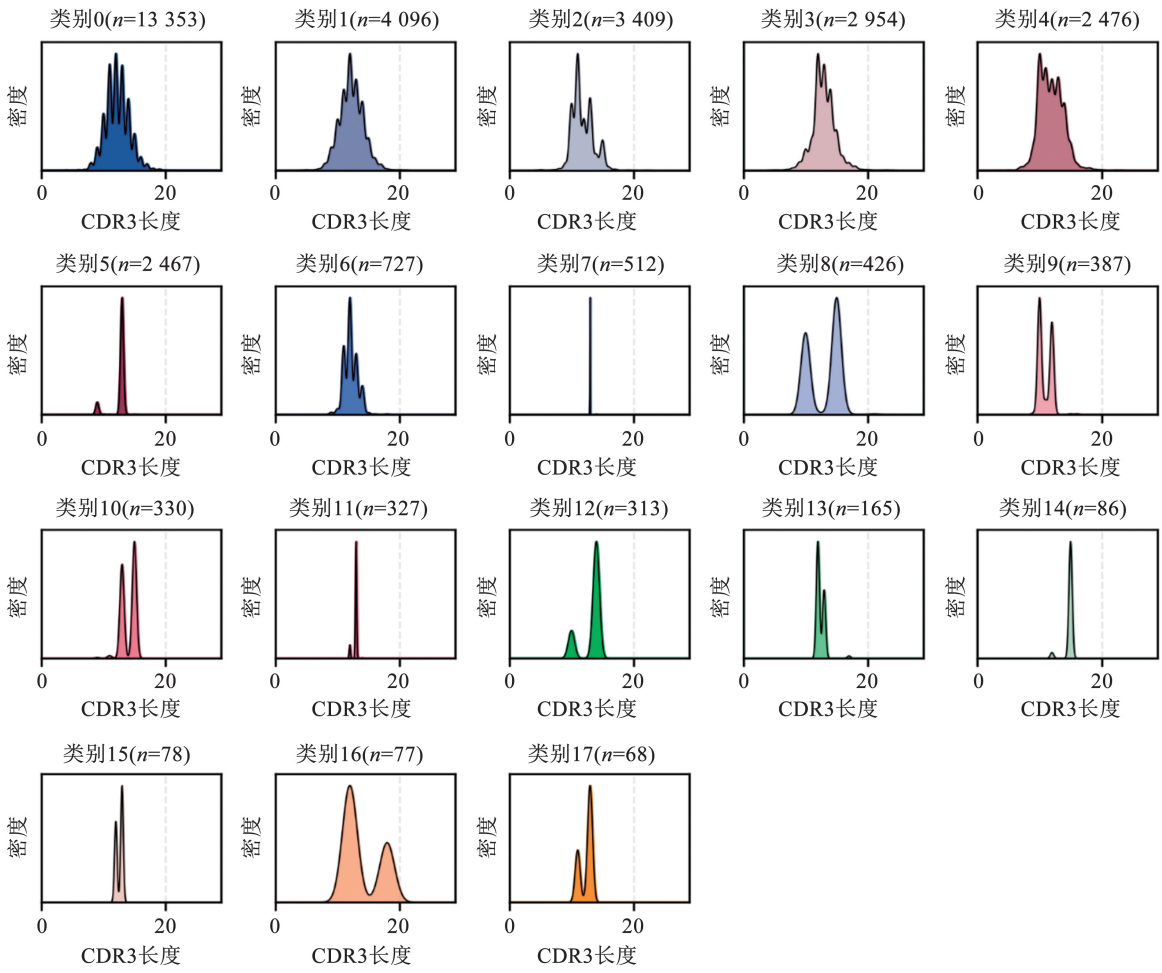
(c) 含量最多的前十位克隆型在整合空间上的分布

图 3 scRTIA 可以在潜在空间中保持克隆型的类别聚集

Fig. 3 scRTIA preserves clonal-type clustering in the latent space



(a) 基于潜在表示的聚类在基因表达空间上的可视化　(b) 基于潜在表示的聚类在潜在空间上的可视化



(c) CDR3序列长度山脊图分布

图 4　scRTIA 可以部分利用免疫特性划分新细胞类别

Fig. 4　scRTIA enables partial segregation of novel cell types using immune-repertoire features

3.2.4　scRTIA 可以检测异质性数据集中的特异性细胞群体

通过前两节发现，scRTIA 可以对单一 T 细胞群体中的免疫信息进行充分提取和整合。对于不同实验，不同条件下得到的生物样本数据集，识别其中具有某些免疫和转录组特性的细胞亚群也具有重要的生物学意义。本研究使用年老脊髓未损伤小鼠、年老脊髓损伤小鼠和年轻脊髓损伤 28 d 后的小鼠3 种不同生物学样本构成的数据集进行分析，其中，每组样本进行了 4 次生物学重复，总共是 12 个不同的数据集。本研究将其组合到一起进行综合分析。

由图 5(a) 可以看出，只利用原始转录组数据，

数据按照其样本分组以及多次生物学重复分成了较分散并且无联系的孤立聚集分布。在进行整合后，如图 5(b)所示，不同条件样本细胞分布更加聚集，并且同一条件下的细胞没有严格按照其生物学重复进行聚集，而是融合了免疫组库信息，其独立出来的分组带有某些免疫特征。

由图 5(c)～(e)中圈出的细胞亚群显示，这些细胞在 T 细胞的 v，，j 基因重排中出现了不同基因型的重排富集，对于出现较多的 TRBV19、TRAV3N-3、TRBV10 等基因在某离散类别中提取出来。

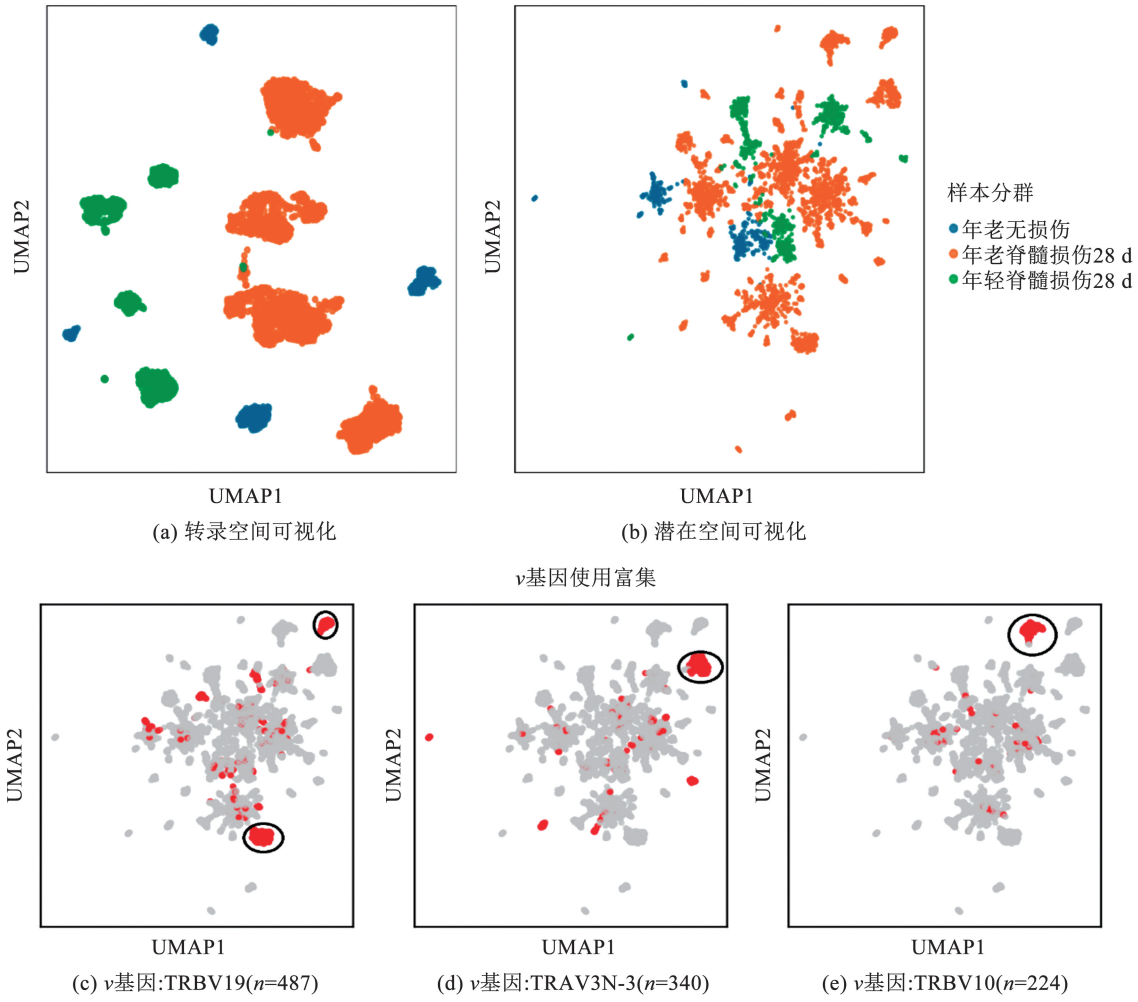


图 5 scRTIA 可以检测异质性数据集中的特异性细胞群体

Fig. 5 scRTIA detection of specific cell populations in heterogeneous datasets

4 结 论

1)采用 Transformer 架构对免疫组库数据进行数值嵌入，在保留关键生物学特征的基础上，使用对 TCR 双链的表示学习。通过基于高斯混合分布的变分自编码器将两条单链信息融合成免疫特征，最终使用概率模型进行双模态整合。

2)在方法构建上，对之前相关研究中将 T 细胞整体看作研究起点的思路进行细化，将分析和整合的起点放在每一个 TCR 单链上，通过对配对的 TCR 单链进行整合得到免疫组表示，可以从更精细的角度对不同的 T 细胞亚群进行分类与刻画。

3)在实验验证上，分别使用同质性和异质性 T 细胞数据集进行方法验证，均得到相关的生物学结论，实现有意义的双模态整合，得到某些既保持转录组特性，又融合免疫特性的 T 细胞亚群。

4)本方法也有一定的局限性，使用时需要注意数据输入的格式。免疫组格式应该具有完整的配对双链信息，而转录组数据应具有. h5 格式或者. mtx 格式，其他格式数据应进行格式转换之后进行输入。本方法在数据预处理方面具有一定的局限性，选择删除了全部的非配对细胞，对于数据的充分使用上具有一定可提升空间，此外，本方法未针对性地处理多数据集之间的批次效应，对于部分多样本数据集

的处理有一定的局限性。

参考文献

[1] CHEN G, NING B, SHI T. Single-cell RNA-seq technologies and related computational data analysis[J]. Frontiers in Genetics, 2019, 10: 317. DOI: 10.3389/fgene.2019.00317

[2] BAYSOY A, BAI Z, SATIJA R, et al. The technological landscape and applications of single-cell multi-omics [J]. Nature Reviews Molecular Cell Biology, 2023, 24(10): 695. DOI: 10.1038/s41580-023-00615-w

[3] IRAC S E, SOON M S F, BORCHERDING N, et al. Single-cell immune repertoire analysis[J]. Nature Methods, 2024, 21(5): 777. DOI: 10.1038/s41592-024-02243-4

[4] SCHATTGEN S A, GUION K, CRAWFORD J C, et al. Integrating T cell receptor sequences and transcriptional profiles byclonotype neighbor graph analysis (CoNGA)[J]. Nature Biotechnology, 2022, 40(1): 54. DOI: 10.1038/s41587-021-00989-2

[5] ZHANG Z, XIONG D, WANG X, et al. Mapping the functional landscape of T cell receptor repertoires by single-T cell transcriptomics[J]. Nature Methods, 2021, 18(1): 92. DOI: 10.1038/s41592-020-01020-3

[6] ZHU B, WANG Y, KU L T, et al. scNAT: a deep learning method for integrating paired single-cell RNA and T cell receptor sequencing profiles[J]. Genome Biology, 2023, 24(1): 292. DOI: 10.1186/s13059-023-03129-y

[7] DROST F, AN Y, BONAFONTE-PARDAS I, et al. Multi-modal generative modeling for joint analysis of single-cell T cell receptor and gene expression data[J]. Nature Communications, 2024, 15(1): 5577. DOI: 10.1038/s41467-024-49806-9

[8] LAI W, LI Y, LUO O J. MIST: an interpretable and flexible deep learning framework for single-T cell transcriptome and receptor analysis[J]. Science Advances, 2025, 11(14): eadr7134. DOI: 10.1126/sciadv.adr7134

[9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all youneed[J]. Advances In Neural Information Processing Systems, 2017, 30: 6000. DOI:10.48550/arXiv.1706.03762

[10] WONG W K, LEEM J, DEANE C M. Comparative analysis of the CDR loops of antigen receptors[J]. Frontiers in Immunology, 2019, 10: 2454. DOI: 10.3389/fimmu.2019.02454

[11] VANDER HEIDEN J A, MARQUEZ S, MARTHANDAN N, et al. AIRR community standardized representations for annotated immune repertoires[J]. Frontiers in Immunology, 2018, 9: 2206. DOI: 10.3389/fimmu.2018.02206

[12] ZHANG B, ROESNER L M, TRAIDL S, et al. Single-cell profiles reveal distinctive immune response in atopic dermatitis in contrast to psoriasis[J]. Allergy, 2023, 78(2): 439. DOI: 10.1111/all.15486

[13] PETREMAND R, CHIFFELLE J, BOBISSE S, et al. Identification of clinically relevant T cell receptors for personalized T cell therapy using combinatorial algorithms[J]. Nature Biotechnology, 2025, 43(3): 323. DOI: 10.1038/s41587-024-02232-0

[14] CRISTALDI V, PATRUNO L, KALLIKOURDIS M, et al. The immune cell dynamics in the peripheral blood of cHL patients receiving anti-PD1 treatment[J]. Frontiers in Oncology, 2025, 15: 1518107. DOI: 10.3389/fonc.2025.1518107

[15] KONG G, SONG Y, YAN Y, et al. Clonally expanded, targetable, natural killer-like NKG7 T cells seed the aged spinal cord to disrupt myeloid-dependent wound healing[J]. Neuron, 2025, 113(5): 684. DOI: 10.1016/j.neuron.2024.12.012

（编辑　刘　彤）