arXiv:2601.00277v1 [q-bio.QM] 1 Jan 2026

单细胞基因组学中的预处理与整合方法基准评测

Ali Anaissi[1,2],Seid Miad Zandavi[2,3], Weidong Huang[1], Junaid Akram[2],Basem Suleiman[4], Ali Braytee[1] 和Jie Hua[5]

[1] 澳大利亚悉尼科技大学[2] 澳大利亚悉尼大学[3] 美国布罗德研究所[4] 澳大利亚新南威尔士大学[5] 中国邵阳学院 ali.anaissi@uts.edu.au, szandavi@broadinstitute.org, weidong.huang@uts.edu.au, Junaid.Akram@uts.edu.au, b.suleiman@unsw.edu.au, ali.braytee@uts.edu.au, steven.hua@mq.edu.au

摘要。单细胞数据分析有望通过在单细胞水平表征疾病相关分子变化，彻底革新个性化医疗。先进的单细胞多模态检测如今能够同时测量数十万个单个细胞中的多种分子（例如 DNA、RNA、蛋白质），从而提供全面的分子读出。一个重要的分析挑战是整合不同模态的单细胞测量。为应对这一挑战，人们开发了多种方法，但尚未针对不同预处理策略对这些技术进行系统评估。本研究考察了单细胞数据分析的通用流程，包括归一化、数据整合和降维。不同算法组合的性能通常取决于数据集规模及其特征。我们使用三项指标——轮廓系数、调整兰德指数和 Calinski-Harabasz 指数——评估了涵盖多种模态、组织和生物体的六个数据集。我们的实验涉及七种归一化方法、四种降维方法和五种整合方法的组合。结果表明，Seurat 和 Harmony 在数据整合方面表现出色，其中 Harmony 的时间效率更高，尤其是在处理大型数据集时。UMAP 是与整合技术兼容性最高的降维方法，而归一化方法的选择则因所用整合方法而异。

1 引言

技术进步显著提升了我们生成高通量单细胞基因表达数据的能力[17]。然而，单细胞数据通常来自多项实验，而实验在采集时间、人员、试剂、设备和技术平台方面存在差异，从而导致较大的变异

，在数据整合过程中可能会混淆生物变异。单细胞 RNA 测序整合[9, 2, 26] 主要应对两个问题：生成细胞类型特征簇，以及判定聚类是否代表真实的细胞类型，还是由生物或技术变异造成，例如特定的批次效应或线粒体含量过高。尽管具有潜力，单细胞 RNA 测序整合仍面临风险，包括因无意义的变异而识别出低质量聚类，以及因相似细胞类型排列不当而产生有偏差的聚类。

Haghverdi 等人提出的一种流行策略 [4] 通过寻找互为最近邻（MNN）[4, 17]，识别数据集之间的细胞映射，并在共享空间中重构数据。该方法虽然能生成适用于下游分析的标准化基因表达矩阵，但计算开销较大。为解决这一问题，fastMNN 算法在通过主成分分析计算得到的子空间中应用 MNN 技术，从而提升性能和准确性[8]。类似地，Scanorama 在降维后的区域中搜索 MNN，以进行批次整合 [6]。

单细胞 RNA 测序整合分析通常包含四个模块：数据标准化、降维、数据整合和结果可视化。每个模块都有大量可用算法，因此可能的组合数量庞大，需要通过评估来确定最佳性能。组合的性能很大程度上取决于数据集的大小和类型，因此难以确定最佳算法和参数设置。应对这一挑战需要大量计算资源、时间和专业知识。

本文通过引入一个实证评估框架来应对这一挑战，帮助科学家评估单细胞 RNA测序算法，并为其数据集选择最佳组合。我们使用多种评估方法，探究适用于不同类型数据集的最优聚类模型组合。该框架分为三个部分：数据标准化、降维和数据整合。在归一化方面，我们研究了七种核心方法：对数归一化、每百万计数（CPM）、SCTransform、词频-逆文档频率（TF-IDF）、Linnorm、Scran 和 TTM [18, 31, 32]。在降维方面，我们评估了主成分分析、UMAP、t-SNE 和 PHATE。在数据整合方面，我们评估了 Seurat、Harmony、FastMNN [4, 17], ComBat [7], 和 Scanorama [6]。我们使用三项评估指标——轮廓系数、调整兰德指数和 Calinski-Harabasz 指数——来考察聚类性能和时间效率。

我们的研究根据各个数据集的评估结果选出最佳模型，并分析不同组合表现各异的原因。我们还深入探讨了针对不同类型和规模数据集的方法选择规律，为未来的模型选择提供数据支持。

我们的主要贡献如下：

1. 我们提出了一个实证框架，用于系统评估单细胞RNA测序数据整合的各种计算策略。该框架涵盖七种归一化方法、四种降维技术和五种整合方法，为单细胞RNA测序数据分析提供了全面的方法。

2. 我们采用轮廓系数、调整兰德指数和 Calinski-Harabasz 指数等稳健的评估指标，分析了这些方法的 140 种组合。该评估阐明了它们在性能、效率和可扩展性方面的表现，并为其在细胞类型聚类及不同来源数据集对齐中的适用性提供了重要见解。
3. 我们的比较分析找出了最适用于单细胞RNA测序数据的归一化、降维和整合方法组合。这为研究人员提供了战略路线图，有助于高保真地整合异质单细胞数据集，并深化对生物学的认识。

2 相关工作

单细胞RNA测序（scRNA-seq）改变了细胞表型的发现与表征方式，有助于在生物医学领域识别生物标志物 [30]。正如Tang等人所述，单细胞RNA测序的基本原理是测量细胞群体中的基因表达分布 [24]。自2014年以来，相关进展大幅降低了测序成本、改进了实验方案，并拓宽了其应用范围。单细胞RNA测序在描绘T淋巴细胞的分子调控方面发挥了关键作用，带来了关于分子决定因素的新认识 [1]。人类细胞图谱（HCA）全球联盟利用这项技术绘制人体组织的参考图谱，有望推动人们对衰老、疾病及潜在治疗方法的理解。未来的应用还将延伸至基于细胞的模型、细胞疗法和再生医学。

然而，单细胞RNA测序数据面临诸多挑战，其中尤为突出的是批次效应。数据采集和处理过程中的差异会导致批次效应，从而妨碍数据整合与解读。Seurat被广泛用于减轻批次效应并整合各种类型的单细胞数据。它采用典型相关分析（CCA）和锚定技术，通过加权最近邻分析处理基因表达差异 [5]。尽管Seurat很有用，但当批次数量较多时，其性能可能会下降，尤其是在处理非高变基因时 [13]。为解决这一问题，Lakkis等人提出了CarDEC，这是一种深度学习模型，通过去噪来提升单细胞RNA测序数据的信息含量。Peng等人 [21] 提出了cFIT方法，这是一种无监督方法，能够以更少的限制整合来自多个来源的数据，从而改善批次效应校正。

归一化对于减少批次效应，同时保留生物变异至关重要 [3]。TMM 等技术已取得成功，但可能过度校正，因此有人建议采用专为单细胞RNA测序设计的方法，如Linnorm 和 SCnorm [18]。单细胞RNA测序数据具有高维度、稀疏性和噪声等特点，通常需要通过降维将其转换到较低维度的空间，同时保留有意义的特性。主成分分析、UMAP、t-SNE 和深度计数自编码器（DCA）等方法各有优缺点；UMAP 能保留全局结构，但也可能引入噪声 [27]。UMAP 和 t-SNE 等可视化方法

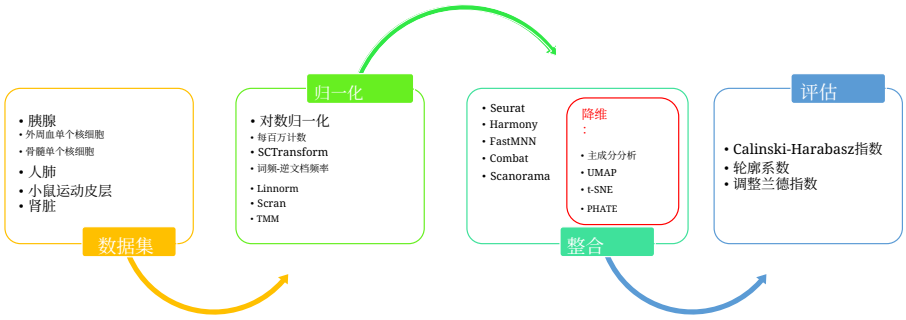


图 1：方法流程

直观易用，可用于评估整合效果，但仍需借助局部逆辛普森指数、平均轮廓宽度和调整兰德指数等定量指标进行严谨评估 [15]。

3 方法论

我们提出了一个用于整合单细胞RNA测序数据的综合框架，包含多个阶段：数据预处理、降维、数据整合以及聚类性能评估。每个阶段均采用多种成熟方法，以确保结果稳健且准确。

首先，数据预处理包括使用多种方法进行归一化，包括对数归一化、每百万计数（CPM）、SCTransform、词频-逆文档频率（TF-IDF）、Linnorm、Scran 和 M 值截尾均值（TMM）。归一化后，采用主成分分析（PCA）、统一流形近似与投影（UMAP）、t分布随机邻域嵌入（t-SNE）和基于亲和力转移嵌入的热扩散势（PHATE）等降维技术，将高维数据转换到较低维度的空间中，以便进行可视化检查和进一步分析。

鉴于各类降维方法在分离生物学簇和检测稀有细胞群体方面的表现各不相同，我们系统地评估了这些方法与不同单细胞 RNA 测序整合方法结合使用时的效果。

接下来，我们使用 Seurat、Harmony、快速互为最近邻（FastMNN）、Combat 和 Scanorama 等常用方法整合处理后的数据并对细胞进行聚类。我们使用降维图展示结果，并通过轮廓系数、Calinski-Harabasz指数和调整兰德指数评估聚类性能。图1展示了我们提出的框架。

3.1 归一化方法

我们选择了以下归一化方法，作为框架的一部分进行研究：

– **对数归一化**：此方法使用对数函数将较大的数值缩放到较小的区间，从而降低较大数值权重的影响，提高模型准确性。– **每百万计数（CPM）**：CPM 将计数列除以其片段总数，再按百万进行缩放，随后进行对数转换。Stuart 等人 [23] 使用此方法在降维前对单细胞 ATAC 测序基因矩阵进行缩放和过滤。– **SCTransform**：SCTransform 是一种用于归一化和方差稳定化的算法，它使用正则化负二项模型，以测序深度为解释变量、UMI 计数为响应变量，为每个基因构建广义线性模型 [3]。– **词频-逆文档频率**：作为文本分析中的标准方法，TF-IDF 根据基因（词语）在细胞（文档）中的频率和逆文档频率，分析基因的重要性 [20]。– **Linnorm**：这种归一化方法使用线性模型和正态性，对单细胞RNA测序数据集进行准确的统计和分析，并以严格筛选出的同源基因为参照 [29]。– **Scran**：Scran 是用于 RNA 测序数据分析的 R 包，其 computeSumFactors 方法通过反卷积校正细胞特异性偏差 [12]。– **M 值截尾均值（TMM）**：TMM 使用对数表达比的加权截尾均值来估算 RNA 产量，并通过计算 M 和 A 值对数据进行归一化；这两个值分别表示对数表达比和平均表达水平。

3.2 降维方法

以下降维方法作为我们所提出框架的一部分进行了研究：

– **主成分分析**：一种线性降维方法，主成分分析会将相关变量转换为少量互不相关的主成分 [16]。– **统一流形逼近与投影**：一种非线性降维技术，与其他方法相比，它能更好地保留数据的全局结构，并具有出色的运行性能 [22]。– **t-SNE**：该方法将高维数据转换到低维空间，同时保持降维前后数据点的概率分布。t-SNE 在低维空间中使用 t 分布，以提升聚类之间的区分度 [10]。– **PHATE**：一种用于高维数据的可视化方法，PHATE 能保留数据的全局结构，并呈现数据点之间的信息几何距离。它对噪声具有鲁棒性，且可扩展至大型数据集 [19]。

3.3 整合方法

整合方法对于去除不需要的技术变异，同时保留有效的生物变异至关重要。我们采用以下整合方法进行批次校正和数据整合：

– **Seurat**：一个为单细胞转录组测序与分析设计的 R 包。Seurat 可整合多种类型的单细胞数据，并分析单细胞转录组测量中的异质性。– **Harmony**：一种用于整合大型单细胞数据集的高效算法。Harmony 首先在低维嵌入空间中对细胞进行聚类，然后根据一种会惩罚不恰当聚类组成的度量，迭代优化这些聚类 [11]。– **FastMNN**：该方法采用改进的互为最近邻（MNN）方法校正批次效应，在降维后识别 MNN 配对，并据此校正批次效应 [33]。– **ComBat**：作为一种经验贝叶斯框架，ComBat 通过标准化数据、估计批次效应参数，并根据这些估计值调整数据来校正批次效应 [7]。– **Scanorama**：该方法通过全景批次校正与整合，整合来自不同技术的单细胞数据集。它采用奇异值分解进行降维，并构建最近邻图以实现整合 [6]。

3.4 数据分析

数据分析采用Wilcoxon秩和检验。这种非参数检验用于比较两个独立样本的分布，以确定它们是否来自同一分布。计算每个数据集的轮廓系数、Calinski-Harabasz指数和调整兰德指数得分后，我们对各方法进行排名，并应用Wilcoxon秩和检验，以确定最佳的归一化、降维和整合方法 [14]。

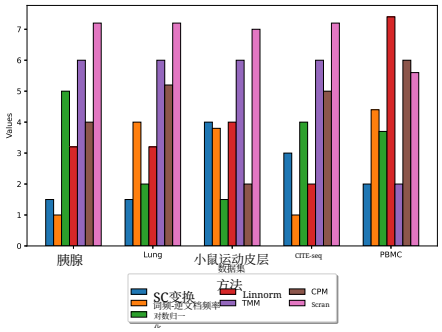
通过遵循这一全面的方法，我们旨在系统评估各种归一化、降维和整合方法在单细胞RNA测序数据分析中的表现，确保在不同数据集上获得稳健且准确的结果。

4 实验与结果

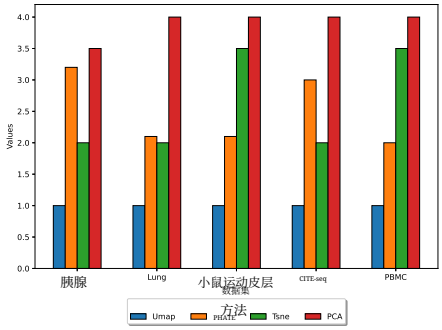
本节讨论五个不同数据集上的模型性能评估和时间效率评估。

4.1 数据集

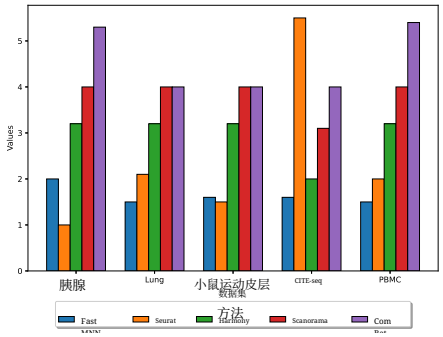
我们在五个 RNA基因序列数据集上进行了实验，具体描述如下：



（a）归一化方法的Wilcoxon秩和检验。



（b）维度重-duction方法。



（c）整合方法的Wilcoxon秩和检验。

Fig. 2：不同方法的 Wilcoxon 秩和检验

表1：评估结果汇总

方法	CH 总和	轮廓系数总和	ARI 总和	CH 排名	轮廓系数排名	ARI 排名	总排名	最终排名
汇总归一化结果								
对数归一化	995503.848	32.3981	63.158689	5	7	7	19	1
CPM	904642.351	28.8685	58.4329301	3	3	3	9	5
SCTransform	1130895.78	31.8076	60.8940582	7	6	6	19	1
同频-逆文档频率	1004035.56	31.4413	60.7435297	6	5	5	16	3
Linnorm	935391.89	29.7849	59.721815	4	4	4	12	4
Scran	743686.811	9.28917	40.5276838	1	1	1	3	7
TMM	813891.279	27.62767	58.2278707	2	2	2	6	6
降维结果汇总								
PCA	203086.58	26.3973	83.709358	1	1	1	3	4
UMAP	2694239.70	67.8274	112.5972185	4	4	4	12	1
TSNE	1205768.39	56.3184	101.296	2	3	3	8	2
PHATE	2380788.68	39.88214	94.727	3	2	2	7	3
整合结果摘要								
Seurat	1144219.94	47.57594	85.112	2	4	4	10	2
Harmony	1525235.92	39.9727	82.208	4	3	3	10	2
FastMNN	1538792.95	50.8561	96.145	5	5	5	15	1
Combat	846507.87	21.849	59.1045765	1	1	1	3	5
Scanorama	1251230.84	30.9551	79.137	3	2	2	7	4

– **胰腺** 数据集由来自八项研究、采用五种不同技术的多个胰腺单细胞RNA测序数据集组合而成。它使用 Seurat[23]整合为单细胞数据集。

– **外周血单个核细胞（PBMC）**数据集基于参与HIV疫苗试验的八名志愿者生成。它采集了疫苗接种后第 0、

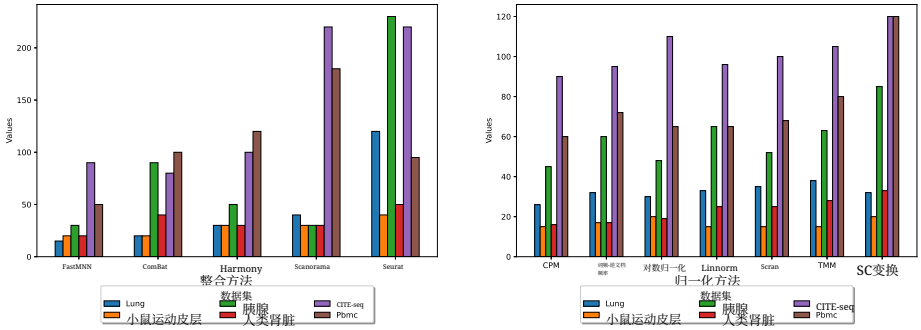
3 和 7 天的三个时间点样本，共形成 24 个样本，并通过 CITE-seq技术处理，生成 RNA 和抗体衍生标签（ADT）。– **CITE-seq** 数据集包含来自八名供体的 30,672 个骨髓单个核细胞（BMNC）样本和 25 种抗体。由人类细胞图谱生成的 BMNC数据集包含 RNA 和抗体衍生标签（ADT）两种检测数据。– **人肺细胞**数据集 [25] 包含来自 65,662 个人肺和血细胞的 58 种分子细胞类型，包括支气管、细支气管、肺泡和循环血液。– **小鼠运动皮层**数据集参考了Yao等人[28] 的研究，该研究分析了成年小鼠新皮层和海马结构，并从 12 只小鼠中构建了转录组和表观基因组图谱。

4.2 方法性能评估

根据各数据集的评估结果，确定了多种表现最佳的方法组合。为汇总所有指标的排名，我们采用 Wilcoxon 秩和法，根据 CH、SC 和 ARI 指标分别对方法进行排名。秩和分数越低，表示在不同数据集中计算脊线高度时表现越好。根据六个数据集上的排名分数总和，方法由上至下排列，表现最佳的方法位于顶部。此外，横轴上的数据集按数据集大小升序排列，数据集大小根据各样本中的特征数计算。

就归一化方法而言（图 2a），SCTransform、对数归一化和词频-逆文档频率是整体表现最突出的前三种方法。这些方法在四个数据集（包括胰腺、外周血单个核细胞、CITE-seq 和小鼠运动皮层）中均排名前三。SCTransform 在胰腺和 CITE-seq 数据集上获得了最高质量的归一化结果，但在小鼠运动皮层数据集上的结果较差。除肺数据集外，对数归一化在所有数据集中均跻身前四。此外，词频-逆文档频率在小鼠运动皮层和肺数据集中得分最高，在外周血单个核细胞数据集中排名前三。词频-逆文档频率和 SCTransform 在处理小型数据集时表现良好，而 SCTransform 也能成功处理较大的数据集。图 2a 显示，对数归一化在大型数据集上的表现更好，且随数据集增大呈下降趋势。

通常，批次整合通过检查 t-SNE 或 UMAP 图进行视觉评估，而我们的实验还使用了 PHATE 图。图 2b 展示了 UMAP 相较于其他降维方法的显著优势。在不限制数据集大小的情况下，UMAP 在所有数据集中始终排名第一，这证明 UMAP 对单细胞RNA测序整合的降维过程具有积极作用。在总体评估指标结果中，PHATE 位居第二，并且其评估表现在大多数数据集中都十分突出。TSNE 和主成分分析是表现最差的方法



（a）不同数据集间整合方法的比较。（b）不同数据集间归一化方法的比较。图3：运行效率排名

尤其是在降维方面，主成分分析在全部五个数据集中都是效果最差的方法。

用于评估不同数据集相关整合方法的评价指标见第4节，其中详细说明了各方法的排名。如图2c所示，根据计算得到的秩和，FastMNN排名第一，Seurat和Harmony并列第二（见表1）。FastMNN在小鼠运动皮层、胰腺和PBMC数据集上表现最佳，但在CITE-seq上的表现较差。与其他方法相比，Scanorama方法的效果最差。由此可以得出结论：FastMNN方法适合处理任何规模的数据集。总体而言，Seurat在较小的数据集上表现更好，而Harmony在大规模和小规模数据集上均表现良好。

4.3 时间效率评估

计算时间是评估模型优劣的另一个重要因素。图 3 展示了整合算法和标准化方法在运行效率方面的比较分析。由于不同数据集的文件大小、运行环境和硬件设备条件存在显著差异，为避免干扰，仅比较同一数据集上不同方法之间的差异。对于数据整合方法而言，尽管 Seurat 的聚类性能最佳，但 Seurat 方法耗时较长。总体而言，耗时最短的方法是 FastMNN，且其聚类性能也相对较好，这意味着该方法更为理想。

对于数据归一化方法，不同方法之间没有太大差异，但 SCTransform 方法耗时较长。不过，考虑到其出色的性能，这段较长的耗时是可以接受的。此外，在分析模型的聚类性能时，Linnorm 和 SCTransform 的表现几乎同样出色；但就效率而言，Linnorm 具有显著优势，因此更适合作为标准化方法。在

总而言之，针对不同数据集的最优模型需要综合确定。以上讨论仅供参考。

5 结论

我们对由不同预处理方法和整合方法组成的工作流程进行了比较分析，以评估其在六个数据集上的表现。结果表明，有必要根据不同数据集的规模及其他特征选择不同的工作流程。此外，对于大型数据集，使用其子集可以显著提高比较不同整合方法的效率。我们基于七种归一化方法、四种降维方法和五种整合方法的组合开展了实验。结果表明，在数据整合模块中，Seurat 和 Harmony 的聚类性能较为突出，但 Harmony 的时间效率更高。同时，Seurat 在小型数据集上的表现更优。在降维模块中，UMAP 方法与整合方法的兼容性表现出良好的前景。由于 FastMNN 的计算时间显著较短，因此建议优先尝试该方法，其他方法也可作为可行的替代方案。

参考文献

1. Aldridge, S., Teichmann, S.A.：单细胞转录组学走向成熟。《自然·通讯》**11**(1)，4307（2020） 2. Alyassine, W., Raju, A.S., Braytee, A., Anaissi, A., Naji, M.：使用变分自编码器实现高效且可靠的单细胞RNA测序数据插补方法。载于：计算研究创新国际会议。第84–97页。施普林格（2024） 3. Hafemeister, C., Satija, R.：使用正则化负二项回归对单细胞RNA测序数据进行归一化和方差稳定化。《基因组生物学》**20**(1)，296（2019） 4. Haghverdi, L., Lun, A.T., Morgan, M.D., Marioni, J.C.：通过匹配互为最近邻校正单细胞RNA测序数据中的批次效应。《自然·生物技术》**36**(5)，421–427（2018） 5. Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W.M., Zheng, S., Butler, A., Lee, M.J., Wilk, A.J., Darby, C., Zager, M. 等：多模态单细胞数据的整合分析。《细胞》**184**(13)，3573–3587（2021） 6. Hie, B., Bryson, B., Berger, B.：使用Scanorama高效整合异质性单细胞转录组。《自然·生物技术》**37**(6)，685–691（2019） 7. Johnson, W.E., Li, C., Rabinovic, A.：使用经验贝叶斯方法校正微阵列表达数据中的批次效应。《生物统计学》**8**(1)，118–127（2007） 8. Jolliffe, I.T., Cadima, J.：主成分分析：综述与最新进展。《皇家学会哲学汇刊A辑：数学、物理与工程科学》**374**(2065)，20150202（2016） 9. Kharchenko, P.V.：单细胞RNA测序计算方法的成就与局限。《自然·方法》**18**(7)，723–732（2021）

10. Kobak, D., Berens, P.：使用t-SNE进行单细胞转录组学分析的艺术。《自然·通讯》**10**(1)，5416（2019） 11. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.r., Raychaudhuri, S.：使用Harmony快速、灵敏且准确地整合单细胞数据。《自然·方法》**16**(12)，1289–1296（2019） 12. L. Lun, A.T., Bach, K., Marioni, J.C.：跨细胞汇总数据，以归一化含有大量零计数的单细胞RNA测序数据。《基因组生物学》**17**，1–14（2016） 13. Lakkis, J., Wang, D., Zhang, Y., Hu, G., Wang, K., Pan, H., Ungar, L., Reilly, M.P., Li, X., Li, M.：联合深度学习模型实现单细胞转录组学中的批次效应校正、去噪和聚类。《基因组研究》**31**(10)，1753–1766（2021） 14. LaMorte, W.W.：曼-惠特尼U检验（Wilcoxon秩和检验）。波士顿大学公共卫生学院（2017） 15. Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Sol´ıs, D.Y., Duque, R., Bersini, H., Now´e, A.：微阵列表达数据整合的批次效应去除方法：综述。《生物信息学简报》**14**(4)，469–490（2013） 16. Liu, Z.：使用半监督主成分分析可视化单细胞RNA测序数据。《国际分子科学杂志》**21**(16)，5797（2020） 17. Lun, A.：MNN算法的进一步开发。GitHub代码库（2019） 18. Lytal, N., Ran, D., An, L.：单细胞RNA测序数据的归一化方法：一项经验性调查。《遗传学前沿》**11**，501166（2020） 19. Moon, K.R., Van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., Elzen, A.v.d., Hirn, M.J., Coifman, R.R. 等：可视化高维生物数据中的结构与转变。《自然·生物技术》**37**(12)，1482–1492（2019） 20. Moussa, M., M˘andoiu, I.I.：使用基于词频-逆文档频率的方法进行单细胞RNA测序数据聚类。《BMC基因组学》**19**，31–45（2018） 21. Peng, M., Li, Y., Wamsley, B., Wei, Y., Roeder, K.：通过CFIT实现单细胞转录组的整合与迁移学习。《美国国家科学院院刊》**118**(10)，e2024383118（2021） 22. Schofield, F., Lensen, A.：使用遗传编程为UMAP嵌入寻找函数映射。载于：2021 IEEE进化计算大会（CEC）。第704–711页。IEEE（2021） 23. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.：单细胞数据的全面整合。《细胞》**177**(7)，1888–1902（2019） 24. Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A. 等：单细胞的mRNA测序全转录组分析。《自然·方法》**6**(5)，377–382（2009） 25. Vieira Braga, F.A., Kar, G., Berg, M., Carpaij, O.A., Polanski, K., Simon, L.M., Brouwer, S., Gomes, T., Hesse, L., Jiang, J. 等：人肺细胞普查揭示健康状态和哮喘中的新型细胞状态。《自然·医学》**25**(7)，1153–1163（2019） 26. Wu, Y., Ji, Y., Lee, S., Akram, J., Braytee, A., Anaissi, A.：用于癌症组织病理学稳健且可扩展诊断服务的简化群体学习框架。载于：国际计算科学会议。第225–232页。施普林格（2025） 27. Xiang, R., Wang, W., Yang, L., Wang, S., Xu, C., Chen, X.：单细胞RNA测序数据降维方法的比较。《遗传学前沿》**12**，646936（2021）

28. Yao, Z., Liu, H., Xie, F., Fischer, S., Adkins, R.S., Aldridge, A.I., Ament, S.A., Bartlett, A., Behrens, M.M., Van den Berge, K., 等：小鼠初级运动皮层的转录组学与表观基因组学细胞图谱。《自然》**598**(7879)，103–110（2021） 29. Yip, S.H., Wang, P., Kocher, J.P.A., Sham, P.C., Wang, J.：Linnorm：改进单细胞 RNA-seq 表达数据的统计分析。《核酸研究》**45**(22)，e179–e179（2017） 30. Yu, B., Li, L., Zhang, J., Wang, X., Zeng, Y.：单细胞测序与甲基化。施普林格（2020） 31. Zandavi, S.M., Koch, F.C., Vijayan, A., Zanini, F., Mora, F.V., Ortega, D.G., Vafaee, F.：基于功率谱密度的特征提取解析单细胞组学表征。《核酸研究》**50**(10)，5482–5492（2022） 32. Zandavi, S.M., Liu, D., Chung, V., Anaissi, A., Vafaee, F.：Fotomics：基于傅里叶变换的组学图像化，用于利用单细胞组学特征谱进行基于深度学习的细胞身份映射。《人工智能评论》**56**(7)，7263–7278（2023） 33. Zhang, F., Wu, Y., Tian, W.：一种去除单细胞数据批次效应的新方法。《细胞发现》**5**(1)，46（2019）