DOI: 10.11844/cjcb.2024.11.0015
中国细胞生物学学报 Chinese Journal of Cell Biology 2024, 46(11): 1985–1996 CSTR: 32200.14.cjcb.2024.11.0015

scRNA-seq与scATAC-seq数据整合分析方法及其在生物医学中的应用

孟子行[1] 李艳国[2] 戎浩[3] 廖奇[3]*
([1]宁波大学医学部公共卫生学院, 宁波 315211;
[2]宁波大学新药技术研究院, 宁波 315211; [3]宁波大学医学部, 宁波 315211)

摘要 随着测序技术的发展, 单细胞转录组测序(single-cell RNA sequencing, scRNA-seq)与单细胞染色质可及性测序(single-cell assay for transposase-accessible chromatin sequencing, scATAC-seq)的出现可以在单细胞层面探索基因表达与染色质结构的开放程度, 而两种测序数据的整合分析也为生物医学研究带来了开辟性的分析思路和方法。研究者可以获得更为全面的细胞功能与细胞状态的信息、细胞内复杂的基因调控网络, 从而揭示生命活动的本质和规律。该文全面探讨了scRNA-seq与scATAC-seq数据整合工具的分析流程、原理及特点, 并展示了这些工具在疾病机理研究、肿瘤学、发育生物学等领域的应用成果, 凸显了整合分析技术在揭示细胞分子机制、理解疾病进程及开发新治疗策略中的独特价值。

关键词 单细胞转录组测序; 单细胞染色质可及性测序; 分析工具; 整合分析; 基因调控网络

The Integration of scRNA-seq and scATAC-seq Data Analysis Methods and Their Applications in Biomedicine

MENG Zixing[1], LI Yanguo[2], RONG Hao[3], LIAO Qi[3]*
([1]*School of Public Health, Health Science Center, Ningbo University, Ningbo 315211, China*; [2]*Institute of Drug Discovery Technology, Ningbo University, Ningbo 315211, China*; [3]*Health Science Center, Ningbo University, Ningbo 315211, China*)

Abstract With the advancement of sequencing technologies, the advent of scRNA-seq (single-cell RNA sequencing) and scATAC-seq (single-cell assay for transposase-accessible chromatin sequencing) allows researchers to explore gene expression and chromatin accessibility at the single-cell level. The integrative analysis of these two types of sequencing data provides groundbreaking analytical approaches and methodologies for biomedical research. Researchers can obtain a more comprehensive understanding of cellular functions and states, as well as the intricate gene regulatory networks within cells, thereby uncovering the fundamental principles and mechanisms of life. This manuscript provides a comprehensive discussion of the analytical processes, principles, and characteristics of available tools for integrating scRNA-seq and scATAC-seq data. It also demonstrates their application outcomes in areas such as disease mechanisms, oncology, and developmental biology. The unique value of integrative analysis techniques is emphasized in revealing cellular molecular mechanisms, understanding disease progression, and developing novel therapeutic strategies.

收稿日期: 2024-07-17 接受日期: 2024-10-08
宁波市自然科学基金(批准号: 2021J124)和宁波市重点研发计划暨“揭榜挂帅”项目(批准号: 2023Z226)资助的课题
*通信作者。Tel: 15857425243, E-mail: liaoqi@nbu.edu.cn
Received: July 17, 2024 Accepted: October 8, 2024
This work was supported by the Ningbo Municipal Natural Science Foundation (Grant No.2021J124) and the Ningbo Key Research and Development Plan and “Reveal the List” Project (Grant No.2023Z226)
*Corresponding author. Tel: +86-15857425243, E-mail: liaoqi@nbu.edu.cn

Keywords single-cell RNA sequencing; single-cell assay for transposase-accessible chromatin sequencing; analytical tools; integrative analysis; gene regulatory networks

整合分析技术的出现在数据的整理与发掘方面为我们提供了强有力的帮助[1-2]，而这一切都得益于测序技术的不断发展，从最初的第一代测序技术(Sanger sequencing)[3]，SANGER团队通过“链终止法”实现了对DNA序列的解析，为基因组学研究奠定了基础；到第二代测序技术(next-generation sequencing, NGS)[4]的兴起，标志着高通量测序的时代到来，大大加快了测序速度，降低了测序成本，使得大规模基因组研究成为可能；再到当前的第三代高通量测序(third-generation sequencing, TGS)[5]，进一步提升了测序的长度和精度，能够直接读取单个DNA分子的序列信息。这些技术手段的不断更迭无不为解读遗传信息、解析基因功能、阐明分子调控机制、了解生命体各种生理活动的机理提供了重要的依据。随着测序技术的进步，单细胞测序技术应运而生，包括单细胞基因组测序(single-cell DNA sequencing, scDNA-seq)[6]、单细胞蛋白质组测序[7]、单细胞染色质可及性测序(single-cell assay for transposase-accessible chromatin sequencing, scATAC-seq)[8]、单细胞代谢组学测序[9]等。这些技术允许从单个细胞中提取和分析遗传物质、蛋白质、代谢物，揭示了细胞间的微小差异和复杂的细胞群体动态，为生物医学研究开辟了全新的视野。

单细胞转录组测序(single-cell RNA sequencing, scRNA-seq)从单一细胞中提取mRNA，并将其逆转录为cDNA。然后，对转录后的cDNA进行修饰和扩增，进而进行高通量测序[10]。目前scRNA-seq已在众多生物过程如神经系统发育、早期胚胎发育以及不同组织和疾病状态的研究中得到广泛应用[11]。与传统的转录组测序技术相比，scRNA-seq能够揭示不同细胞类型之间的转录组差异，发现罕见的细胞类型，并且可以识别细胞在不同生理状态下的基因表达模式[12]。scATAC-seq使用Tn5转座酶对单个细胞的基因组DNA进行标记，可以特异性切割开放染色质区域，即那些未被核小体紧密包裹的DNA区域，切割后的DNA片段被测序，从而获得染色质可及性信息[13]。这些信息允许我们识别DNA序列，包括启动子、增强子、绝缘子、沉默子等顺式调控元件，通过揭示这些调控元件的位置和活性，可以增强我们对基因表达调控的理解，为疾病研究和新治疗方法的开发提供新的视角。

scRNA-seq主要用于分析单个细胞的基因表达情况，但它无法提供例如启动子、增强子活性等基因调控的信息，而scATAC-seq虽然能揭示染色质结构和调控元件的开放性，但无法直接告诉我们哪些基因正在活跃表达。两者结合可以更加全面了解基因表达的调控机制与细胞不同状态下独特的功能模式，这种结合分析手段也是识别疾病关键调控区域和基因的重要手段，目前已在众多生物过程中被广泛应用。如美国宾夕法尼亚大学的FASOLINO等[14]研究者通过结合scRNA-seq和scATAC-seq来研究1型糖尿病(type 1 diabetes, T1D)患者中人类胰岛的细胞状态与功能变化，并进一步分析了T1D胰岛细胞的基因表达和调控机制。该研究发现T1D患者的胰管细胞表现出与树突状细胞相似的转录特征，这表明胰管细胞可能参与了T1D的免疫反应，为理解T1D的免疫病理机制提供了新的视角，并为开发潜在的治疗方法奠定了基础；美国哈佛大学BUENROSTRO等[15]研究者利用scATAC-seq技术，对10种人类造血细胞类型进行了详细的分析，研究发现了共同髓系前体细胞(common myeloid progenitors, CMPs)和粒细胞−巨噬细胞前体细胞(granulocyte-monocyte progenitors, GMPs)内部存在显著的异质性。同时，通过结合scRNA-seq数据，研究发现，在造血细胞不同的分化阶段中，GATA2、PU.1、RUNX1等转录因子的活性会发生变化，并且它们能够通过调整染色质可及性来控制关键基因*MYC*和*TAL1*的表达。通过这些分析，研究者们揭示了如启动子和增强子之类的调控元素与它们潜在目标基因的直接联系，为理解造血过程中血液细胞转变的分子机制提供了新的思路。

scRNA-seq和scATAC-seq数据的整合分析，综合考虑了基因表达模式和调控模式，更加精准地解析了复杂生物过程中的调控机制，这些数据分析依赖于不断涌现的生物信息学工具，包括scDART[16]、SeuratV3[17]、GLUE[18]等(表1)。本文就关于这些工具的原理、分析流程、应用及优缺点作一系统

表1 scRNA-seq与scATAC-seq数据整合工具

Table 1 Tools for integrating scRNA-seq and scATAC-seq data

工具名称 Tool name	算法原理 Algorithm principle	功能 Functionality	兼容性及集成 Compatibility and integration	参考文献 References
scDART (single cell deep learning model for ATAC-seq and RNA-seq trajectory integration)	Deep learning framework, learning a low-dimensional latent space to preserve the cellular trajectory structure	Handles batch effects and accurately depicts gene activity	Designed for integration with TensorFlow and PyTorch	[16]
Seurat v3	CCA (canonical correlation analysis)	Integrates highly variable features in scRNA-seq and scATAC-seq data and maps the data to a shared space	Supports R environment	[19]
LIGER (linked inference of genomic experimental relationships)	iNMF (integrative non-negative matrix factorization)	Integrates scRNA-seq and scATAC-seq data, revealing common biological processes across datasets	Supports R environment	[20]
scNCL (single-cell neighborhood contrastive learning)	Converts scATAC-seq features into a gene activity matrix and introduces neighborhood contrastive learning	Through neighborhood contrastive learning, it preserves the neighborhood structure of the original feature space; learns transferable latent features with feature projection loss and alignment loss for high-accuracy label transfer	Supports Python environment	[21]
scAWMV	AWMV (adaptively weighted multi-view) learning, matrix decomposition, and neighborhood contrastive learning	Generates biologically meaningful low-dimensional representations, reveals latent structures of cell states, and improves the accuracy of cell heterogeneity analysis	Supports Python environment	[22]
scBridge	Heterogeneous transfer learning	Evaluates the accuracy of cell classification through reliability modeling, optimizes integration quality, and effectively enhances data fusion precision and efficiency	Supports Python environment	[23]
GLUE (graph-linked unified embedding)	VAE (variational autoencoder), guidance graph, adversarial learning	Embeds information into a shared low-dimensional space, enabling cross-omics data integration and gene regulatory network inference	Supports Python environment	[18]
MOFA (multi-omics factor analysis)	Factor analysis and VI (variational inference)	Identifies latent factors across different datasets and maps them into a unified latent factor matrix	Supports R and Python environments	[24]
scEMC (single-cell effective multi-modal clustering)	SAN (skip aggregation network) and ZINB (zero-inflated negative binomial) denoising autoencoder	Integrates scRNA-seq and scATAC-seq data to improve clustering performance while retaining rich information	Supports Python environment	[25]
scJoint (single-cell joint learning)	Pre-trained neural network, semi-supervised learning	Performs label transfer and clustering optimization, supports large-scale data processing	Supports Python environment	[26]
MAESTRO (model-based analyses of single-cell transcriptomics and regular omics)	Collaborative (NMF non-negative matrix factorization)	Projects data into a shared feature space to reveal cell types and regulatory networks	Supports Python environmen	[27]
SMGR (single-cell multi-omics gene co-regulatory)	ZINB (zero-inflated negative binomial) and generalized linear regression model	Identifies consistently expressed genes and peaks, determines co-regulatory programs, and uncovers gene regulatory networks specific to cell types	Supports Python environmen	[28]

续表 1

工具名称 Tool name	算法原理 Algorithm principle	功能 Functionality	兼容性及集成 Compatibility and integration	参考文献 References
sciCAN (single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network)	Cycle-consistent adversarial network	Uses an encoder to project high-dimensional chromatin accessibility and gene expression data into a shared low-dimensional space, effectively representing the biological information of both data types	Supports Python environmen	[29]
coupleCoC+	Clusters and matches similar features, transferring knowledge from source datasets (e.g., scRNA-seq data) to assist in analyzing target datasets (e.g., scATAC-seq or sc-methylation data)	Utilizes features from the target data that are not directly linked to the source data, improving data analysis accuracy and efficiency by clustering multiple layers of source and target data simultaneously	Supports Python environmen	[30]
Con-AAE (contrastive cycle adversarial autoencoders)	Adversarial autoencoder and self-supervised contrastive learning	Projects different modalities of data (e.g., scRNA-seq and scATAC-seq) into a low-dimensional manifold space, enhancing consistency between different data modalities	Supports Python environmen	[31]

综述, 为生物医学研究者更好地选择分析工具提供依据。

1 scRNA-seq和scATAC-seq数据整合分析框架

通过对scRNA-seq和scATAC-seq数据的整合分析, 能够揭示出单独使用任意一种数据所无法发现的生物学洞见。近年来已经有许多的研究者针对scRNA-seq与scATAC-seq的数据分析进行综述, LUECKEN等[32]对scRNA-seq数据, 从上游数据处理, 到下游的可视化呈现, 详尽细致地整理了scRNA-seq数据分析的全部流程, 同时评估了现有的分析工具。与此同时, LAREAU等[33]的综述工作很好地描述评估了针对scATAC-seq数据的分析工具, 这些研究都为后续的研究者们提供了很好的参考。在数据整合分析方面, MA等[34]全面评估了单细胞多模态组学(single-cell multimodal omics, scMulti-omics)技术的发展与挑战, 重点讨论了如何在单个细胞中同时测量多种生物模态信息, 并对这些数据加以整合的方法, 此外, 文中还综述了多个用于scMulti-omics分析的计算工具, 并通过案例研究展示了它们的实际应用效果, 为研究人员进行工具选择时提供了一个全面的参考框架。

目前, scRNA-seq与scATAC-seq数据的整合分析思路主要包括以下几点: 首先, 选择细胞间合适的锚点(anchor); 其次, 构建综合特征矩阵与细胞类型注释, 使研究者能够在同一分析框架下观察并分析这两种数据, 揭示转录活动与染色质状态之间的相互作用; 最后, 在数据融合与映射的基础上, 研究者可以进一步构建基因间的共表达网络和染色质区域的共可及性网络, 来揭示基因表达调控和染色质状态调控之间的联系, 及其在生理和疾病条件下的关键作用(图1)。“锚点”选择(anchor identification)是scRNA-seq与scATAC-seq数据整合分析的第一步, 也是最为关键的一步。在实际应用中, 锚点选择的方法取决于输入数据的类型, 同时, 不同数据模态之间的匹配程度(即是否存在一一对应的细胞或特征)也是锚点选择的重要限制因素。根据锚点的不同选择, 单细胞数据整合方法可以分为水平整合、垂直整合和对角整合三类。水平整合通常用于将同类型但来自不同批次或技术平台的数据进行整合, 在水平整合中, 常用的锚点选择方法主要依赖于基因表达的共性或细胞的邻近关系, 常用的方法包括MNN(mutual nearest neighbors)匹配、LIGER、Harmony等; 垂直整合用于从同一细胞中同时获取不同模态的数据, 正是因为这一特点, 在垂直整合中, 通

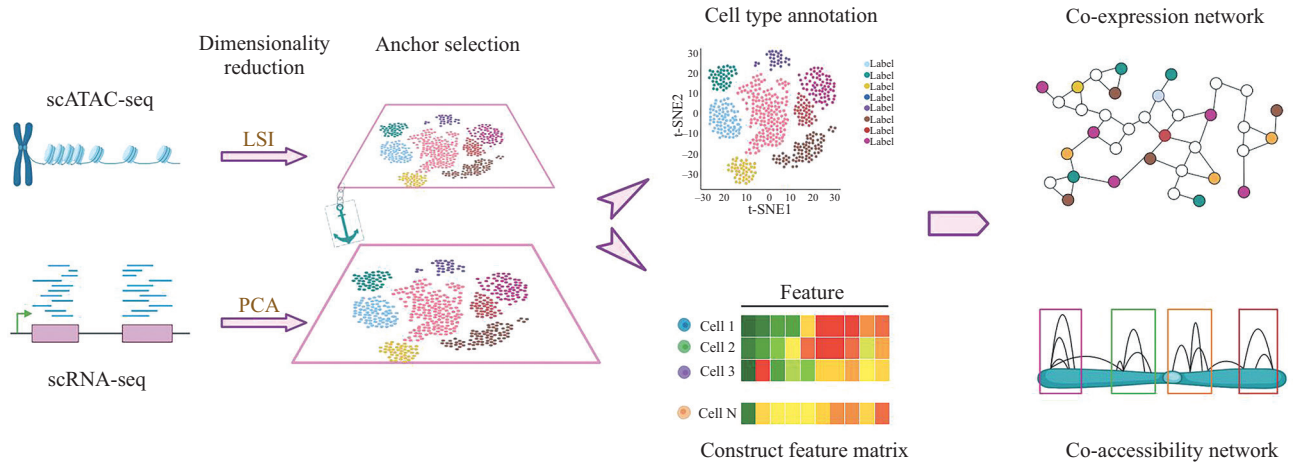

图1 scRNA-seq与scATAC-seq数据整合流程

Fig.1 Workflow for integrating scRNA-seq and scATAC-seq data

常使用细胞作为整合的锚点, 常用的方法包括直接匹配或使用MOFA(multi-omics factor analysis)将多模态数据映射到一个潜在的共享空间中, 从而通过细胞作为锚点进行整合; 对角整合用于数据模态和细胞群都不相同的实验设计中, 因此很难找到直接匹配的锚点, 它是通过复杂的数学模型来推断不同模态之间的潜在协作关系, 从而实现数据的整合[1]。由于scRNA-seq与scATAC-seq数据所使用的测序手段不同, 所包含的基因组特征信息也不同, 无法在高维空间找到合适的锚点, 因此在寻找锚点之前需要使用主成分分析(principal component analysis, PCA)、基于变分自编码器(variational auto-encoder, VAE)以及LSI(latent semantic indexing)等方法对二者数据进行降维处理; 在降维后的数据中, 可以使用MNN匹配[12]、基于共调控区域的匹配[35]、典型相关分析(canonical correlation analysis, CCA)以及加权最近邻(weighted nearest neighbor, WNN)分析[36]等方法寻找二者数据之间的锚点, 这个步骤确保了两种数据类型中相似细胞群体的一一对应, 为后续整合分析奠定了基础。

在找到锚点后, 通常采用加权平均法来对scATAC-seq数据的矩阵进行修正, 使其与scRNA-seq数据的结构更为一致。接下来将修正后的scATAC-seq数据与scRNA-seq数据整合到一个矩阵中, 每一行代表一个细胞, 每一列代表一个特征(如基因或染色质区域)。整合后的矩阵通常仍然是高维的, 因此要进一步利用机器学习和统计方法, 如PCA、t分布随机邻域嵌入(t-distributed stochastic neighbor embedding, t-SNE)或者一致流形近似和投影(uniform manifold approximation and projection, UMAP), 对矩阵进行降维处理, 并最终得到一个包含每个细胞的基因表达水平和染色质开放区域的综合特征矩阵。

细胞类型注释, 首先通过聚类分析初步识别出各个细胞群体, 然后, 使用已知的标记基因和染色质可及性签名对这些细胞簇进行注释, 来确定它们的细胞类型, 这一步结合了scRNA-seq表达谱和scATAC-seq调控区特征, 可以更加精准地识别细胞类型; 功能注释与生物学解释, 这一步骤是将数据分析的结果转化为对生物学机制的理解。在这一步骤中, 可以应用如基因本体论(Gene Ontology, GO)或京都基因与基因组百科全书(Kyoto Encyclopedia of Genes and Genomes, KEGG)路径分析等方法, 识别出与特定细胞类型或细胞状态相关的生物过程与功能通路。总的来说, 数据整合与映射确保了研究者在后续分析过程中, 能够同时兼顾基因表达与染色质状态的信息。这一步骤在整体分析流程中发挥着至关重要的作用。

在数据整合后, 可以进一步分别构建基因间的共表达网络与染色质区域的共可及性网络[37]。共表达网络反映基因间的表达相似性, 可以识别出在特定细胞群体中共同上调或下调的基因, 从而推断可能参与相同生物过程或由相同转录因子调控的基因关系[31]。类似地, 共可及性网络识别在不同细胞类型或状态中共同可及的染色质区域, 可以揭示潜在的转录因子结合位点和调控元件之间的相互作用, 从而为理解基因的表达调控提供深入见解。整合这

两个网络, 研究者可以继续探索基因表达模式与染色质开放区域之间的相关性, 并识别其中可能存在的调控关系。不同于初步的融合整理, 这一阶段旨在揭示更复杂的调控机制, 如特定的转录因子结合位点与基因表达水平之间的定量关系, 或者染色质可及性如何影响基因的调控网络。此外, 研究者还可以进一步探索基因表达调控中的时空特异性, 识别在特定发育阶段或细胞应答中活跃的调控元件。这些分析有助于揭示细胞类型特异性的调控机制, 为理解生物学复杂性和疾病机理提供更全面的视角。

在得到了特征矩阵、细胞类型、功能注释、共表达网络与共可及网络的基础上, 研究者还可以进一步深入分析特定基因表达模式背后的染色质可及性, 推断关键调控元件(如增强子和启动子)的活性。这一方法通过线性回归、逻辑回归和生存分析等统计模型以及随机森林、支持向量机、神经网络等机器学习算法, 对细胞分化、增殖或对治疗的响应和基因调控进行更深入的理解和预测。如通过对健康与疾病状态的比较, 揭示关键的调控路径和转录因子等。这种方法进一步提升了对调控元件活性推断的准确性, 使研究人员能够揭示不同细胞状态下调控元件的动态变化, 更加深入地理解基因调控网络及其在疾病中的关键作用。

2 scRNA-seq与scATAC-seq数据的整合分析工具

2.1 Seurat v3

Seurat v3是2019年由STUART等[20]开发的基于R语言的分析包, 为scRNA-seq与scATAC-seq数据的整合分析贡献了一整套从质量控制、数据标准化、聚类和差异表达分析的完整分析流程。Seurat v3通过引入“锚点”(anchors)的概念, 识别并连接不同单细胞数据集中的细胞, 这些“锚点”是由两个来自不同数据集的细胞组成的, 虽然这些细胞的来源不同, 但它们在生物学特征上是相同或相似的, 通过找到这种对应关系(即锚点), Seurat v3可以将不同数据集有效地整合在一起。在整合scRNA-seq和scATACseq数据时, Seurat v3的关键步骤是通过scATAC-seq的peak matrix生成基因活性矩阵(gene activity matrix)的, 这一过程至关重要。这个过程的假设和方法建立在基因调控与染色质可及性之间的关系之上, 我们通常认为基因的启动子区域和其他调控区域的开放状态与其表达水平存在相关性, 因此, scATAC-seq数据中峰的强度可以用来推断基因的活性, 基于这种假设, Seurat v3可以构建一个可以反映基因活性的矩阵来代表基因表达, 基于这个矩阵scATAC-seq的数据能够与scRNA-seq数据在一个共同的框架中进行整合和分析。在方法上, Seurat v3会根据每个基因的启动子或调控区域(如增强子)内scATAC-seq峰的强度估算每个基因在细胞中的活性, 进而将scATAC-seq数据转化为一个类似于基因表达矩阵的表征。为了更好地处理两个数据集之间的技术差异, Seurat v3使用正则化的CCA[38]将基因活性矩阵和scRNA-seq的表达矩阵投射到一个共享的低维空间中, 并通过L2归一化进一步优化数据对齐, 在该共享空间中, Seurat v3利用MNN的算法[12], 识别两种数据集中生物学上相似的细胞对, 这些细胞对被称为“锚点”。识别出锚点后, Seurat v3会对每个锚点进行评分, 这些分数是基于两个数据集中相关细胞邻域结构的一致性来计算的。也就是说, 如果某个锚点所在的细胞在各自数据集中都具有类似的邻居(即在生物学上相似的细胞), 那么这个锚点的分数就会更高, 这一过程使得不同来源的数据集能够在整合后更加准确地反映出它们的生物学特性, 从而进行更可靠的下游分析。

相对于其他工具而言, Seurat v3精确的锚点识别与评分过程, 使得其能够在整合数据时有效地消除技术噪声, 保持生物学信号的一致性。并且Seurat v3的架构使其能够灵活应对不同规模的数据集, 从小型数据集到包含数百万个细胞的大型数据集, 无论是整合两个数据集还是多个数据集, Seurat v3都能保持高效、准确的整合效果。正是因为这样的特性, Seurat v3成为当前scRNA-seq与scATAC-seq数据整合分析最常用的工具之一, 被广泛应用。

2.2 GLUE

GLUE(graph-linked unified embedding)是由北京大学高歌等[18]研究人员开发, 用于整合分析不同组学单细胞数据的算法框架。GLUE算法结构包括以下几个关键部分: 首先, GLUE使用变分自编码器(variational autoencoder, VAE)的机器学习模型, 从每个组学层(如scRNA-seq或scATAC-seq数据)中学习到一个更加简单、低维的表示方法, 来帮助研究者更好地理解并整合来自不同组学层的数据。其

次, GLUE使用了一种叫做“指导图”(guidance graph)的工具, 该图包含了不同组学数据(比如基因表达数据和染色质开放性数据)中的重要特征, 这些特征可以是基因、染色质区域等。在这个“指导图”中, 每个特征被表示为一个“顶点”或“节点”, 而“边”则表示这些特征之间的相互关系或调控作用。举个例子, 如果我们知道某个基因的表达可能受到它附近的染色质开放性区域的调控, 那么在这个指导图中, 这个基因和染色质区域之间会有一条“边”连接, 这条边表示它们之间的调控关系。这个图的作用是帮助GLUE将来自不同组学数据层的信息整合在一起, 让这些数据能够相互关联和对齐, 最终帮助研究者更好地理解这些数据背后的生物学意义。最后, GLUE采用对抗性学习方法, 将不同组学层的细胞嵌入在共享的低维空间中对齐, 在这个过程中“指导图”帮助模型确定哪些特征之间的关系应该在不同组学层中保持一致, 从而确保最终的嵌入结果能够正确反映跨层之间的生物学联系。

与其他工具相比, GLUE的优点在于其高精度、强鲁棒性、优异的可扩展性和灵活的模块化设计。高精度体现在GLUE能够非常准确地将不同组学层的数据对齐, 能够正确地反映出不同组学数据中相同的生物学信息。强鲁棒性体现在GLUE可以在“指导图”有一定误差的情况下, 依旧能够找到合理的嵌入表示, 并确保最终数据对齐的效果不受太大影响。优异的可扩展性体现在GLUE使用了神经网络架构和小批量优化技术, 使得它能够在数据量非常庞大(数百万个细胞)的情况下, 仍然以较低的计算成本进行高效分析。最后, GLUE的模块化设计赋予了其极大的灵活性, 能够轻松适应不同的研究需求, 甚至可以实现跨物种的数据整合, 为研究复杂单细胞多组学数据提供了强有力的支持。

2.3 LIGER

美国哈佛大学WELCH等[20]开发的LIGER工具, 不同于Signac、ArchR专为scATAC-seq数据开发的特性, LIGER可以高效地整合来自不同个体、时间点、物种, 以及来自不同分子模式下的测序数据[39]。在数据处理中, LIGER可以将scRNA-seq、scATAC-seq、单细胞蛋白质组等不同类型的单细胞数据集作为输入数据, 进而使用整合非负矩阵分解(integrate nonnegative matrix factorization, iNMF)算法来解析多种单细胞数据集中的复杂生物学信号, 如细胞类型的共同表达模式、特定细胞状态的标记基因, 以及细胞分化或发育过程中的动态变化。iNMF通过在共享的低维空间内对齐相似的细胞状态和细胞类型, 实现了不同数据集间的有效整合, 同时保留了每个数据集特有的生物学信号, 提高了细胞类型识别的准确性, 且允许研究者揭示不同数据集中的共有特征和差异性生物标记, 为复杂生物学现象的解析提供了强大的工具。除此之外, LIGER在数据可视化方面也具有丰富的功能, 提供了如t-SNE和UMAP等多种工具, 这些降维和可视化的技术使得研究者能够直观地观察和分析数据中细胞类型的分类与生物学过程的变化。

2.4 scBridge

scBridge(single-cell Bridge for RNA-seq and ATAC-seq data integration)是由四川大学和杭州电子科技大学的研究团队设计的一种深度学习框架[23], 该框架可充分利用细胞异质性整合不同批次的scRNA-seq和scATAC-seq数据。该框架包括两个主要模块: 深度神经网络编码器和分类器模块, 以及可靠性建模和跨组学对齐模块。首先, 使用已标注的scRNA-seq数据对深度神经网络进行预训练, 赋予模型初步的特征提取和细胞分类能力。接着, 将训练好的网络应用于scATAC-seq数据, 利用高斯混合模型(gaussian mixture model, GMM)评估每个细胞的分类可信度, 识别出那些与scRNA-seq数据差异较小的scATAC-seq细胞(在基因表达和染色质开放状态之间具有较高正相关性的细胞)。随后, 计算并对齐scRNA-seq和scATAC-seq数据中各细胞类型的特征中心(代表该类型细胞的平均特征或“原型”), 逐步缩小组学之间的差异。通过多次迭代, 将最终得出的scATAC-seq细胞加入到标注数据集中, 重复训练过程, 逐步整合更多的scATAC-seq细胞, 不断优化模型性能, 直到所有数据成功整合。开发者还将scBridge在多个数据集(如SNARE-seq、SHARE-seq和10× Multi-ome)上进行测试, 发现scBridge在数据整合和标签转移准确率方面均表现出色, 并显著优于其他基线方法。这项工作不仅展示了深度学习框架在连接不同单细胞技术方面的潜力, 也为后续研究者提供了一个整合scRNA-seq与scATAC-seq数据的强大工具。

2.5 MOFA

MOFA(multi-omics factor analysis)算法, 是ARGELAGUET等[24]开发, 用于整合多组学数据集(如

scRNA-seq、scATAC-seq、DNA甲基化、蛋白质组学等)的算法框架。MOFA首先利用名为因子分析(factor analysis)的统计学技术[40], 对每个组学数据集进行分解, 并从中提取一组潜在因子(latent factors), 进而使用线性模型来表示观测数据与潜在因子之间的关系。具体来说, 每个潜在因子会对最终数据产生一定的“影响”, 这个“影响”用一个权重(或系数)来表示。然后, 将这些潜在因子乘以各自的权重, 再将它们相加在一起, 就可以得到最终的观测数据。这种线性模型帮助我们简化和理解数据之间的复杂关系, 并且通过分析权重的大小, 可以推断出哪些潜在因子对观测数据有更大的影响, 从而揭示出数据背后的生物学机制。在这个过程中, MOFA还使用了一种叫做变分推断(variational inference, VI)的技术[41]来估计每个潜在因子的得分, 通过不断调整和优化, 变分推断帮助MOFA逐步找出每个数据集最可能的潜在因子, 以及这些因子如何影响不同的数据。在确定了潜在因子之后, MOFA就能将所有的数据集映射到同一个“潜在因子空间”中, 并构建一个统一的潜在因子矩阵, 通过这种方式可以将不同的数据类型(如基因表达和染色质数据)用一组相同的潜在因子解释。MOFA的优点在于它通过共享的潜在因子模型, 实现了不同组学数据的整合分析, 相比于其他方法, MOFA通过变分推断来有效处理高维数据和部分缺失数据, 使得整合更加高效, 结果更加稳定。在美国俄亥俄州大学的研究中[34], 研究者使用了MOFA、LIGER和Seurat v3三种整合方法对一组相同的scRNA-seq与scATAC-seq数据集进行测试, 发现MOFA在分析匹配数据集上得分最高, 是三种常用方法中效率最高的整合方法。在MOFA的基础上, 研究者们在2020年进一步开发了MOFA+[42]。与MOFA相比, MOFA+采用了更先进的概率框架, 能够更加自然地处理缺失值, 从而提升了模型在实际应用中的鲁棒性, 在数据不完整时MOFA+仍然能够提供可靠的分析结果, 同时引入了随机变分推断(stochastic variational inference, SVI), 通过GPU加速显著提升了计算速度, 在处理大规模数据集时, MOFA+实现了高达20倍的速度提升, 使其适用于包含数十万细胞的复杂数据集。

scRNA-seq和scATAC-seq数据的整合分析不仅可以增强我们对细胞分子层面的理解, 还推动了基础生物学研究和疾病机理研究的发展, 近年来各种分析工具的出现也为研究者提供了丰富的分析思路。在上述内容中, 我们详细介绍了5种可以用于scATAC-seq与scRNA-seq整合分析的生物信息学工具, 并详细列举了各自的技术功能与优劣之处, 为后续研究者对于不同工具的选择提供了参考, 并且, 随着这些技术工具的不断完善和应用, scRNA-seq和scATAC-seq数据的整合分析技术也将在生物医学研究和临床诊断中发挥更大的作用。

3 scRNA-seq与scATAC-seq整合分析的应用

随着scRNA-seq和scATAC-seq整合分析工具的不断完善与发展, 近年来众多研究者也将这些技术手段应用于不同的领域, 并产出了许多卓越的研究成果(表3)。这些应用不仅限于基础生物学, 还涵盖了肿瘤学等疾病机理研究、药物开发, 以及发育生物学等多个领域, 为我们提供了深入了解生命科学的新窗口。

3.1 肿瘤学领域

在肿瘤学领域, scRNA-seq和scATAC-seq数据的整合分析有助于揭示肿瘤微环境内的复杂互动关系。肿瘤微环境是由多种不同类型的细胞(包括肿瘤细胞、免疫细胞、间质细胞等)组成。通过分析这些细胞的基因表达和染色质状态, 可以更好地理解它们之间的相互作用, 以及这些相互作用如何促进或抑制肿瘤的生长和扩散。

在广西医科大学的一项研究中[43], 研究者利用Seurat、Signac工具, 对19个肾透明细胞癌(clear cell renal cell carcinoma, ccRCC)样本的scRNA-seq与scATAC-seq数据进行整合分析。通过建立ccRCC单细胞转录组和染色质可及性图谱, 揭示了不同肿瘤细胞亚型的调控特性, 并发现了两个能够促进ccRCC侵袭和迁移的长非编码RNA(RP11-661C8.2和CTB-164N12.1)。这项研究强调了在肿瘤微环境中细胞间复杂交互的重要性, 并识别出了促进肿瘤发展的关键lncRNA, 为未来的精准医疗和个体化治疗提供了可能性。在另一项同样针对ccRCC的研究中[44], 研究者利用Seurat工具, 将来自scRNA-seq和scATAC-seq的数据进行了整合分析。首先, 他们对单细胞数据进行聚类和细胞亚群鉴定; 然后利用scATAC-seq数据分析染色质可及性, 并将其与scRNA-seq数据进行关联, 以揭示转录调控网络。最

后结合细胞亚群信息和染色质状态, 研究团队绘制了ccRCC肿瘤微环境的细胞相互作用图谱, 同时, 研究者还进行了$CD8^+$ T细胞和巨噬细胞染色质可及性和基因表达的综合分析, 并揭示了它们亚群中的不同调控元素, 以及肿瘤微环境内由配体–受体相互作用介导的细胞间通信, 进一步解释了ccRCC的细胞异质性, 也为该癌症的预后与治疗提供了关键依据。

3.2 疾病机理研究领域

在疾病机理研究中, scRNA-seq和scATAC-seq数据的整合分析也已经被用来揭示多种疾病(包括癌症、自身免疫疾病、神经退行性疾病等)的细胞层面机制。通过整合分析, 研究者能够识别出特定的基因调控网络, 以及这些网络在疾病中的发生、发展和药物抗性中发挥的关键作用。这些发现不仅为理解不同疾病的生物学基础提供了新的视角, 也为开发更有效的治疗策略提供了潜在的靶点。如在美国斯坦福大学GRANJA等[45]的研究中, 研究者使用Cicero将scATAC-seq数据转换为推断的基因活性得分。然后通过Seurat v3的CCA方法, 将这些基因活性得分与scRNA-seq表达数据对齐到共同空间, 建立了正常血液发育的表观遗传基线, 以解析混合表型急性白血病(mixed phenotype acute leukemia, MPAL)患者血液中的异常分子特征。通过对转录组和染色质可及性的综合分析, 研究者确定了91 601个潜在的峰值−基因联系, 并鉴定了调控白血病特异性基因的转录因子。例如, 他们发现了一个与标志物基因*CD69*邻近的调控元件, 该元件由转录因子RUNX1连接。此研究进一步深入理解了MPAL在不同患者中发现的共同恶性特征和个体特定的调控特征, 为白血病研究提供了新的视角和潜在的治疗靶标。在一项针对皮肤病的研究中, 耶鲁大学BIELECKI等[46]研究者通过Signac与Seurat工具, 对scRNA-seq与scATAC-seq数据进行融合分析, 探讨了皮肤驻留的先天淋巴细胞(innate lymphoid cells, ILCs)在银屑病模型中的作用。研究者发现, 在由IL-23或伊米昆莫德诱导的银屑病中, 皮肤ILCs转变为具有病理性的ILC3样状态, 并且, 研究者揭示了这些细胞在疾病状态下的动态变化, 包括从静息状态到ILC2效应状态的转换。在疾病发展过程中, 这些细胞进一步转变为同时具有ILC2和ILC3特征的混合状态。这一发现强调了皮肤ILCs对外部刺激的响应范围和灵活性, 并提出了健康组织中的免疫活动在未受控制时可能导致病理性改变, 对理解皮肤免疫反应的复杂性和动态性至关重要, 为银屑病的治疗提供新的靶点。

德国海德堡大学的POOS团队[47]整合了全基因组测序、scRNA-seq和scATAC-seq, 以及线粒体DNA突变数据, 来探究15名复发或难治性多发性骨髓瘤患者的亚克隆结构和进化。研究发现, 多发性骨髓瘤亚克隆之间存在显著的表观遗传和转录异质性, 这直接影响了不同亚克隆对治疗的响应和抗药性表现。特别是, 关键的肿瘤促进因子如MYC、BCL-2和NF-κB在一些亚克隆中由于染色质区域更加开放, 导致这些因子表达水平升高, 这一现象表明这些因子是针对难治性疾病的潜在治疗靶点。此外, 研究还发现了亚克隆之间的免疫逃逸机制, 例如通过PD-L1抑制T细胞的活性, 研究还表明, CXCR4及其配体SDF-1的阻断可能有助于抑制骨髓瘤细胞的迁移及其与微环境的相互作用。这些发现为开发新的免疫和细胞微环境靶向治疗策略提供了新的方向。在另一项针对女性绝经后输卵管的研究中[48], 研究者使用了Seurat和Signac工具来整合和分析scRNA-seq与scATAC-seq数据。通过Seurat进行细胞类型的标注和数据整合, Signac用于处理scATAC-seq数据, 评估基因活性分数和进行转录因子活性分析。基于整合结果, 研究者揭示了输卵管和卵巢中基质细胞的主导作用, 这些基质细胞表达了与细胞衰老相关的基因, 如ECM组件和信号通路相关基因。输卵管上皮细胞中则表达了多个与卵巢癌风险相关的基因, 如*BRCA1*、*BRCA2*和*TP53*。此外, 上皮细胞和基质细胞之间存在活跃的细胞通信, 涉及TGF-β、WNT和Hedgehog等信号通路, 特别是输卵管伞部显示出与其他区域不同的基因表达调控特征。这些发现有助于深化我们对绝经后输卵管和卵巢细胞组成的理解, 推动对绝经期妇科疾病的研究和潜在治疗策略的发展。

3.3 发育生物学领域

在发育生物学领域, scRNA-seq和scATAC-seq的结合应用为研究细胞分化和器官发育提供了独特的视角。通过追踪单细胞层面上的基因表达模式和染色质状态的变化, 研究者能够描绘出细胞分化的详细路线图, 对理解复杂生物过程和组织形成至关重要。

如日本横滨RIKEN综合医学科学中心MIYAO

等[49]将整合分析技术应用于髓质胸腺上皮细胞(medullary thymic epithelial cells, mTECs)的研究, 他们使用Signac和Seurat工具, 结合了scRNA-seq数据和scATAC-seq数据, 深入探究了胸腺上皮细胞(thymus epithelial cell, TEC)的异质性和分化动态。研究者通过从染色质可及性区域预测基因表达情况, 并将这些预测与scRNA-seq数据进行对比, 识别出了一种特定的增殖状态胸腺上皮细胞。这些细胞展现出独特的染色质结构, 并高表达自身免疫调节因子Aire及共刺激分子CD80。经过一段时间的扩增后, 这些细胞转变为静息状态, 此时它们高度表达组织特异性抗原(tissue-specific antigens, TSAs)。该研究通过分析这些细胞在分化过程中的暂时性扩增状态, 揭示了胸腺中自我耐受诱导的关键阶段, 并可能为未来的免疫治疗提供新的策略。

在针对血液系统的研究中, 英国剑桥大学RANZONI[50]利用人类胎儿肝脏和骨髓血细胞的scRNA-seq和scATAC-seq数据, 详细分析了从人类胎儿肝脏和骨髓中提取的超过8 000个免疫表型造血干细胞前体(hematopoietic stem and progenitor cells, HSPCs)。研究识别出三个关键的多能前体细胞群体: 多线性原始造血前体细胞(multipotent progenitor cells, MPPs)、早期淋巴造血前体细胞(early lymphoid progenitors, ELPs)和造血干细胞(hematopoietic stem cells, HSCs)。这些细胞展示了向T细胞、B细胞、NK细胞、单核细胞和巨核细胞等多种血细胞系的分化潜能。通过pseudotime轨迹分析, 研究展示了从MPPs到髓系和淋巴系细胞的详细分化路径, 揭示了转录因子如GATA1、PU.1和TCF7在这一过程中的关键调控作用。这些发现为理解胎儿期造血和相关血液疾病提供了重要的分子和细胞基础, 有助于未来血液疾病的研究和治疗策略的开发。

3.4 其他领域

scRNA-seq与scATAC-seq数据的整合分析不仅在上述领域成果丰硕, 在其他领域的研究中也得到了广泛应用。在YOSHIMURA等[51]的研究中, 研究者收集了肾脏类器官分化过程中第7天到第26天的样本, 生成了scRNA-seq与scATAC-seq数据, 并使用Seurat v4工具对其进行了整合。通过整合技术手段, 研究重构了肾单位上皮细胞的发育轨迹, 揭示了肾单位前体细胞逐步分化为不同成熟肾脏细胞类型(如近端小管、亨利袢和集合管)的过程, 并确定了每个分化阶段的关键基因表达谱和染色质开放区域。此外, 研究者进一步通过CRISPR干扰技术, 靶向抑制了HNF1B的启动子和增强子区域, 发现HNF1B的抑制显著阻止了这些细胞类型的分化, 导致了类器官中这些细胞类型的数量减少, 这一发现进一步证明了HNF1B驱动近端小管细胞和亨利袢细胞分化上有着十分重要的作用。这项研究不仅为肾脏类器官的发育和成熟提供了全面的图谱, 同时也为进一步优化类器官技术和理解肾脏相关疾病的分子机制提供了重要的数据支持。在深圳暨南大学的一项研究中[52], 研究者整合分析了维持性血液透析(maintenance hemodialysis, MHD)患者中外周血单核细胞(peripheral blood mononuclear cells, PBMC)的scRNA-seq与scATAC-seq数据。研究发现, 透析显著抑制了$CD4^+$ T细胞中的TCR信号转导以及单核细胞中的MHC II通路, 在$CD4^+$ T细胞中, TCR相关基因(如*TRAV4*、*CD45*等)表达量减少; 在单核细胞中, MHC II相关基因(如*HLA-DRB1*、*HLA-DQA1*等)同样表现出下调趋势。同时, 研究还发现透析患者PBMC亚群间的细胞通讯显著减少, 尤其是T细胞和单核细胞之间的TGF-TGFBR信号通路被削弱, 并且透析还改变了一些与T细胞激活和抗原呈递过程密切相关的通路, 如HVEM-BTLA和IL16-CD4信号。这项研究详细揭示了MHD患者中循环免疫细胞的转录和表观遗传特征, 解释了透析患者免疫功能下降的机制, 提出了透析过程中与免疫相关的分子靶点, 为优化透析设备和技术、减少其对免疫系统的负面影响提供了科学依据。

scRNA-seq与scATAC-seq的整合分析不仅加深了我们对不同疾病过程与机理的理解, 还为未来的疾病预防、诊断和治疗开辟了新的路径。随着这些技术的进一步完善和应用, 将有更多创新性的发现展现在我们眼前, 为解决人类面临的健康问题提供更有效的解决方案。这些技术的发展标志着向个性化医疗和精准治疗迈进的重要一步, 有望实现更为个性化、有效和安全的医疗健康管理。在未来, 随着更多跨学科合作的展开, 整合分析技术的潜力将得到进一步挖掘和应用, 为生命科学研究和医学进步贡献力量。

4 总结与展望

通过对scRNA-seq与scATAC-seq数据的整合分

析工具的整理, 我们列举了这些生物信息学工具所拥有的不同技术特性, 以及每种工具最为适配的应用场景; 这些工具从数据预处理、降噪、细胞类型识别到调控网络分析提供了十分全面的功能支持, 并极大地推动了复杂生物学数据分析的发展。在具体的应用中, 这些整合工具不仅帮助研究者更好地探究肿瘤微环境中的相互作用、肿瘤异质性以及治疗抗性机制, 也在发育生物学领域中更加全面地揭示了细胞分化路径和调控机制。此外, 在探索疾病机理方面, 这些工具通过整合不同层面的分子数据, 揭示了多种疾病状态下的关键分子标志物和潜在的治疗靶点。

尽管这些工具已经在生物医学领域的研究中取得了显著的成果, 但在实际应用中依然面对着稀疏数据的处理、算法效率的提升以及数据跨平台兼容性的挑战; 此外, 目前的研究大多集中在特定类型样本以及疾病模型上, 针对不同种群、不同环境条件下的广泛应用和验证相对有限, 这也在一定程度上限制了整合分析技术在更广泛背景下的推广与应用。在未来的发展中, 解决这些限制和不足的关键在于跨学科合作的加强与分析技术的不断创新; 在人工智能和机器学习技术的加持下, 将生物学、计算科学、数学统计等领域的专业知识和技术进行融合, 并加大对于多样性样本和复杂疾病模型的研究投入, 将有助于拓宽整合分析技术的应用范围, 提高其在不同环境和种群中的适应性和准确性, 从而使其更好地服务于生命科学研究和临床应用。

参考文献 (References)

[1] ARGELAGUET R, CUOMO A S E, STEGLE O, et al. Computational principles and challenges in single-cell data integration [J]. Nat Biotechnol, 2021, 39(10): 1202-15.

[2] AIBAR S, GONZALEZ-BLAS C B, MOERMAN T, et al. SCENIC: single-cell regulatory network inference and clustering [J]. Nat Methods, 2017, 14(11): 1083-6.

[3] SANGER F, NICKLEN S, COULSON A R. DNA sequencing with chain-terminating inhibitors [J]. Proc Natl Acad Sci USA, 1977, 74(12): 5463-7.

[4] MARGULIES M, EGHOLM M, ALTMAN W E, et al. Genome sequencing in microfabricated high-density picolitre reactors [J]. Nature, 2005, 437(7057): 376-80.

[5] GOODWIN S, MCPHERSON J D, MCCOMBIE W R. Coming of age: ten years of next-generation sequencing technologies [J]. Nat Rev Genet, 2016, 17(6): 333-51.

[6] EVRONY G D, HINCH A G, LUO C. Applications of single-cell DNA sequencing [J]. Annu Rev Genomics Hum Genet, 2021, 22: 171-97.

[7] KELLY R T. Single-cell proteomics: progress and prospects [J]. Mol Cell Proteomics, 2020, 19(11): 1739-48.

[8] CUSANOVICH D A, DAZA R, ADEY A, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing [J]. Science, 2015, 348(6237): 910-4.

[9] SVATOS A. Single-cell metabolomics comes of age: new developments in mass spectrometry profiling and imaging [J]. Anal Chem, 2011, 83(13): 5037-44.

[10] TANG F, BARBACIORU C, WANG Y, et al. mRNA-Seq whole-transcriptome analysis of a single cell [J]. Nat Methods, 2009, 6(5): 377-82.

[11] STUART T, SATIJA R. Integrative single-cell analysis [J]. Nat Rev Genet, 2019, 20(5): 257-72.

[12] HAGHVERDI L, LUN A T L, MORGAN M D, et al. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors [J]. Nat Biotechnol, 2018, 36(5): 421-7.

[13] BUENROSTRO J D, WU B, LITZENBURGER U M, et al. Single-cell chromatin accessibility reveals principles of regulatory variation [J]. Nature, 2015, 523(7561): 486-90.

[14] FASOLINO M, SCHWARTZ G W, PATIL A R, et al. Single-cell multi-omics analysis of human pancreatic islets reveals novel cellular states in type 1 diabetes [J]. Nat Metab, 2022, 4(2): 284-99.

[15] BUENROSTRO J D, CORCES M R, LAREAU C A, et al. Integrated single-cell analysis maps the continuous regulatory landscape of human hematopoietic differentiation [J]. Cell, 2018, 173(6): 1535-48,e16.

[16] ZHANG Z, YANG C, ZHANG X. scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously [J]. Genome Biol, 2022, 23(1): 139.

[17] STUART T, BUTLER A, HOFFMAN P, et al. Comprehensive integration of single-cell data [J]. Cell, 2019, 177(7): 1888-902,e21.

[18] CAO Z J, GAO G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding [J]. Nat Biotechnol, 2022, 40(10): 1458-66.

[19] SATIJA R, FARRELL J A, GENNERT D, et al. Spatial reconstruction of single-cell gene expression data [J]. Nat Biotechnol, 2015, 33(5): 495-502.

[20] WELCH J D, KOZAREVA V, FERREIRA A, et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity [J]. Cell, 2019, 177(7): 1873-87,e17.

[21] YAN X, ZHENG R, CHEN J, et al. scNCL: transferring labels from scRNA-seq to scATAC-seq data with neighborhood contrastive regularization [J]. Bioinformatics, 2023, 39(8): btad505.

[22] ZENG P, MA Y, LIN Z. scAWMV: an adaptively weighted multi-view learning framework for the integrative analysis of parallel scRNA-seq and scATAC-seq data [J]. Bioinformatics, 2023, 39(1): btac739.

[23] LI Y, ZHANG D, YANG M, et al. scBridge embraces cell heterogeneity in single-cell RNA-seq and ATAC-seq data integration [J]. Nat Commun, 2023, 14(1): 6045.

[24] ARGELAGUET R, VELTEN B, ARNOL D, et al. Multi-omics factor analysis: a framework for unsupervised integration of

multi-omics data sets [J]. Mol Syst Biol, 2018, 14(6): e8124.
[25] HU D, LIANG K, DONG Z, et al. Effective multi-modal clustering method via skip aggregation network for parallel scRNA-seq and scATAC-seq data [J]. Brief Bioinform, 2024, 25(2): bbae102.
[26] LIN Y, WU T Y, WAN S, et al. scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning [J]. Nat Biotechnol, 2022, 40(5): 703-10.
[27] WANG C, SUN D, HUANG X, et al. Integrative analyses of single-cell transcriptome and regulome using MAESTRO [J]. Genome Biol, 2020, 21(1): 198.
[28] SONG Q, ZHU X, JIN L, et al. SMGR: a joint statistical method for integrative analysis of single-cell multi-omics data [J]. NAR Genom Bioinform, 2022, 4(3): lqac056.
[29] XU Y, BEGOLI E, MCCORD R P. sciCAN: single-cell chromatin accessibility and gene expression data integration via cycle-consistent adversarial network [J]. NPJ Syst Biol Appl, 2022, 8(1): 33.
[30] ZENG P, LIN Z. coupleCoC+: an information-theoretic co-clustering-based transfer learning framework for the integrative analysis of single-cell genomic data [J]. PLoS Comput Biol, 2021, 17(6): e1009064.
[31] WANG X, HU Z, YU T, et al. Con-AAE: contrastive cycle adversarial autoencoders for single-cell multi-omics alignment and integration [J]. Bioinformatics, 2023, 39(4): btad162.
[32] LUECKEN M D, THEIS F J. Current best practices in single-cell RNA-seq analysis: a tutorial [J]. Mol Syst Biol, 2019, 15(6): e8746.
[33] CHEN H, LAREAU C, ANDREANI T, et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data [J]. Genome Biol, 2019, 20(1): 241.
[34] MA A, MCDERMAID A, XU J, et al. Integrative methods and practical challenges for single-cell multi-omics [J]. Trends Biotechnol, 2020, 38(9): 1007-22.
[35] CORCES M R, GRANJA J M, SHAMS S, et al. The chromatin accessibility landscape of primary human cancers [J]. Science, 2018, 362(6413): eaav1898.
[36] HAO Y, HAO S, ANDERSEN-NISSEN E, et al. Integrated analysis of multimodal single-cell data [J]. Cell, 2021, 184(13): 3573-87,e29.
[37] PLINER H A, PACKER J S, MCFALINE-FIGUEROA J L, et al. Cicero predicts cis-regulatory dna interactions from single-cell chromatin accessibility data [J]. Mol Cell, 2018, 71(5): 858-71,e8.
[38] BUTLER A, HOFFMAN P, SMIBERT P, et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species [J]. Nat Biotechnol, 2018, 36(5): 411-20.
[39] LIU J, GAO C, SODICOFF J, et al. Jointly defining cell types from multiple single-cell datasets using LIGER [J]. Nat Protoc, 2020, 15(11): 3632-62.
[40] BUNTE K, LEPPAAHO E, SAARINEN I, et al. Sparse group factor analysis for biclustering of multiple data sources [J]. Bioinformatics, 2016, 32(16): 2457-63.
[41] BLEI D M, KUCUKELBIR A, MCAULIFFE J D. Variational inference: a review for statisticians [J]. J Am Stat Assoc, 2017, 112(518): 859-77.
[42] ARGELAGUET R, ARNOL D, BREDIKHIN D, et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data [J]. Genome Biol, 2020, 21(1): 111.
[43] YU Z, LÜ Y, SU C, et al. Integrative single-cell analysis reveals transcriptional and epigenetic regulatory features of clear cell renal cell carcinoma [J]. Cancer Res, 2023, 83(5): 700-19.
[44] LONG Z, SUN C, TANG M, et al. Single-cell multiomics analysis reveals regulatory programs in clear cell renal cell carcinoma [J]. Cell Discov, 2022, 8(1): 68.
[45] GRANJA J M, KLEMM S, MCGINNIS L M, et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia [J]. Nat Biotechnol, 2019, 37(12): 1458-65.
[46] BIELECKI P, RIESENFELD S J, HUTTER J C, et al. Skin-resident innate lymphoid cells converge on a pathogenic effector state [J]. Nature, 2021, 592(7852): 128-32.
[47] POOS A M, PROKOPH N, PRZYBILLA M J, et al. Resolving therapy resistance mechanisms in multiple myeloma by multiomics subclone analysis [J]. Blood, 2023, 142(19): 1633-46.
[48] LENGYEL E, LI Y, WEIGERT M, et al. A molecular atlas of the human postmenopausal fallopian tube and ovary from single-cell RNA and ATAC sequencing [J]. Cell Rep, 2022, 41(12): 111838.
[49] MIYAO T, MIYAUCHI M, KELLY S T, et al. Integrative analysis of scRNA-seq and scATAC-seq revealed transit-amplifying thymic epithelial cells expressing autoimmune regulator [J]. eLife, 2022, 11: e73998.
[50] RANZONI A M, TANGHERLONI A, BEREST I, et al. Integrative single-cell RNA-Seq and ATAC-Seq analysis of human developmental hematopoiesis [J]. Cell Stem Cell, 2021, 28(3): 472-87,e7.
[51] YOSHIMURA Y, MUTO Y, LEDRU N, et al. A single-cell multiomic analysis of kidney organoid differentiation [J]. Proc Natl Acad Sci USA, 2023, 120(20): e2219699120.
[52] WU H, DONG J, YU H, et al. Single-Cell RNA and ATAC sequencing reveal hemodialysis-related immune dysregulation of circulating immune cell subpopulations [J]. Front Immunol, 2022, 13: 878226.