分类号_____ 学校代码_____

密 级_____ 学 号 0202121690

本科毕业论文（设计）

单细胞 ATAC 数据分析中批次效应矫正算法的基准测试

Benchmarking Algorithms for Batch Correcttion in Single-cell ATAC-seq data Analysis

院系名称 生命科学学院

专业名称 生物工程

专业年级 2020 级生物工程

学生姓名 郭佳亿

指导教师 张垲研究员；郝慧芳副教授

2024 年 4 月 5 日

单细胞 ATAC 数据分析中的批次效应矫正算法的基准测试

摘要

【目的】随着高通量测序技术与单细胞技术的发展，在研究基因组的表达调控网络的过程中，单细胞染色质可及性分析（Single cell ATAC-seq）技术逐渐成为研究基因组表达调控网络的主要手段之一，但在大多数 Single cell ATAC 分析中，其核心问题之一即如何消除不同实验批次导致的后续分析中不需要批次效应（Batch Effect），本研究旨在比较四种矫正批次效应的算法 MNN，Harmony，BBKNN，scVI 在处理两个不同单细胞 ATAC-seq 数据分析任务的表现，以评估它们对生物特征的保留和批次效应的移除效果。对其进行初步的基准测试。【内容】本研究首先从 GEO 数据库中下载发表在网上的 7 组大脑单细胞 ATAC-seq 数据的 fastqc 文件并通过 Trimgalore、BWA-MEM、SAMTOOL 等软件对数据进行预处理，将得到的 Anndata 数据根据存在的批次效应分为两组数据集，使用四种算法分别消除由于不同供体导致的批次效应和由于不同的供体与取样部位导致的批次效应，最后用四种指标评估不同算法的表现以及其对生物特征的保留效果，为算法的选择提供参考。【结论】在处理批次效应来源单一的数据集时 Harmony 通常在所有指标方面均表现出色，处理批次效应来源复杂的数据集时，BBKNN 表现更佳，两种基于线性嵌入模型原理的算法 Harmony 和 MNN 算法在不同的数据集中均表现优异，BBKNN 在混合样本方面明显差于其他几种算法，此外，深度学习算法 scVI 也展现出了稳定的表现，通过这样的系统评估和分析，本研究为生物医学领域中单细胞染色质可及性数据中的批次效应处理提供更深入的理解和有效的解决方案，促进相关研究的发展和应用，同时也进一步揭示了深度学习原理在单细胞数据集分析中的应用前景。

关键词：单细胞 ATAC-seq 技术；批次效应矫正；单细胞数据分析；基准测试

Benchmarking Algorithms for Batch Correcttion in Single-cell ATAC-seq data Analysis

Author:Guo Jiayi

Tutor: Zhang Kai; Hao Huifang

Abstract

【Purpose】 With the development of high-throughput sequencing technology and single-cell techniques, single-cell chromatin accessibility analysis (Single cell ATAC-seq) has gradually become one of the main means to study the regulatory networks of genome expression. However, in most Single cell ATAC analyses, one of the core issues is how to eliminate unnecessary batch effects caused by different experimental batches. This study aims to compare four batch effect correction algorithms (MNN, Harmony, BBKNN, scVI) in handling two different single-cell ATAC-seq data analysis tasks, to evaluate their performance in preserving biological features and removing batch effects, and to conduct preliminary benchmark testing. 【Content】 This study first downloads fastqc files of 7 sets of brain single-cell ATAC-seq data published online from the GEO database, and preprocesses the data using software such as Trimgalore, BWA-MEM, SAMTOOL, etc. The obtained Anndata data is then divided into two datasets based on existing batch effects, and four algorithms are used to respectively eliminate batch effects caused by different donors and those caused by different donors and sampling sites. Finally, four metrics are used to evaluate the performance of different algorithms and their preservation effects on biological features, providing references for algorithm selection. 【Conclusion】 When dealing with datasets with a single source of batch effects, Harmony usually performs excellently in all metrics. When handling datasets with complex batch effects sources, BBKNN performs better. The Harmony and MNN algorithms based on the linear embedding model principle perform outstandingly in different datasets. BBKNN is notably inferior to other algorithms in handling mixed samples. Additionally, the deep learning algorithm scVI also shows stable performance. Through such systematic evaluation and analysis, this study provides a deeper understanding and effective solutions for handling batch effects in single-cell chromatin accessibility data in the field of biomedicine, promoting the development and application of related research, and further revealing the application prospects of deep learning principles in single-cell dataset analysis.

Key words: Single cell ATAC-seq; Batch effect remval; Benchmark; Single cell data analysis

1 绪论

1.1 单细胞 ATAC-seq

在人类基因组计划完成后，高通量测序技术进入了高速发展的阶段[1]，随着高通量测序技术的提出，针对于基因组的表达与调控网络的分析，逐渐成为了基因组学领域研究的主流方向，染色质可及性对于基因组的转录激活以及表达调控过程至关重要[2]，目前已经有大量的文章证明了转录因子（TFs）和组蛋白修饰（HMs）以及非编码 RNA 在染色质可及性区域中扮演着重要的角色[3,4]。因此，针对于染色质可及性的研究能够从基因组的水平分析不同状态下的细胞的基因组转录激活过程中的差异。一些方法例如 DNase-seq，MNase-seq 的提出揭示了研究染色质可及性的主要手段[5]，即对激活基因区域进行片段化，然后再进行高通量测序。染色质可及性分析（Assay for Transposase Accessible Chromatin with high-throughput, ATAC）这一方法于 2016 年首次被提出[6]，该方法使用高活性的 Tn5 转座酶对细胞的全基因组进行处理，结合于染色质的开放区域并将 Tn5 转座酶携带的接头序列插入到这些位点中，通过高通量测序的方法对 Tn5 转座酶捕获到的序列进行测序，从而揭示基因转录活跃的区域，由于其能够通过更加简便快速的操作方式得到全基因组的染色质可及性数据，且相对于之前的 DNase-seq，MNase-seq 等方法，ATAC-seq 需要的样本量更小，数据的分析流程更加简便，重复性更好[7]，到今天为止 ATAC-seq 技术已经成为了研究基因组表达与调控网络的主要研究手段之一[8,9]。

长期以来，样本本身存在的异质性以及细胞之间存在的差异困扰着科学家，作为 ATAC-seq 技术的衍生，单细胞 ATAC-seq 技术（Single-cell ATAC-seq）逐渐取代了之前的批量（bulk）测定细胞染色质开放性并计算平均值的方法[10]。由于不同细胞之间存在着不同的转录因子和组蛋白修饰调控，这些影响因素可能会导致不同细胞中转录因子与 DNA 片段的结合发生变化，从而抑制或者激活相关基因的表达。单细胞 ATAC-seq 技术的出现，有望揭示不同细胞转录调控机制背后更深层次的生物学机制[11]。

1.2 单细胞 ATAC-seq 得到的数据结构类型

单细胞 ATAC-seq 处理后得到的数据通常以 Anndata（Annotated data）的数据格式储存，Anndata 是单细胞测序分析的基础数据结构，可以为储存，操作和管理单细胞

数据提供一种便捷高效的方式。Anndata 是一种表格式的数据结构，与 DataFrame 类似，但是 Anndata 更适用于处理高维生物学数据[12]。如图 1.1 所示，anndata 类型的数据结构主要包括四个部分[13]，数据矩阵（X）：这是主要的数据矩阵，其中每一行对应一个观测值（在单细胞 ATAC 数据中，每一行对应不同染色质区域的开放度），每一列对应一个变量或特征（在单细胞 ATAC 数据中，每一列代表一个细胞）。观测值的注释（obs，obsm，obsp）：这些是与观测值相关的注释。在单细胞 ATAC 测序数据中，每个细胞（观测值）可能有额外的元数据，如供体信息。变量的注释（var，varm，varp）：这些是与变量相关的注释。例如，每个细胞可能有额外的元数据，如替代基因符号。非结构化注释（uns）可以用于存储任何其他非结构化元数据。AnnData 对象可以像 DataFrame 一样进行切片，这使得它非常灵活和便于数据分析。

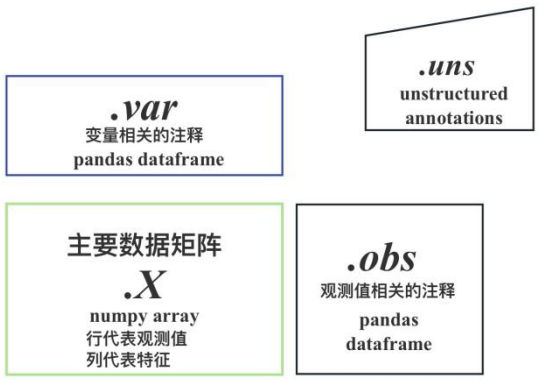


图 1.1 Anndata 数据结构

Figure1.1 The Structure of Anndata

总的来说，Anndata 是一种强大灵活的数据结构，在单细胞数据分析中被广泛使用，在本研究中，我们主要对合并了不同样本的 Anndata 数据集对象进行了批次效应矫正，并对最终的结果进行了可视化分析以及聚类质量的评估。

1.3 单细胞测序数据之间存在的批次效应（Batch Effect）

在大多数 Single cell ATAC 分析中，其核心问题之一即批次效应（Batch Effect）[14]，批次效应是指在不同的批次或组别中处理导致的测得染色质可及性的差异，例如，两个不同实验室如果从同一个批次中采集样本，但这些样本的分离方式不同，则会产生批次效应，一般来说，批次效应的来源是多种多样的，很难确定。一些批次效应可能是由于技术方面的原因导致的，例如测序深度的差异，处理样本手段的不同，或者

是实验方案的不同。一些生物学效应如供体，取样组织的不同，或者是取样位置的不同导致的结果的差异也通常被解释为批次效应[15]。

是否需要消除批次效应在 Single cell ATAC-seq 中的影响取决于实验方案的设计和提出的问题，消除批次效应在联合分析（Joint Analysis）至关重要，只有消除批次效应才能使我们在联合分析中找到不同批次的共同数据结构[16]。也只有在消除批次效应之后才能找到此前由于批次之间的差异被掩盖的稀有细胞群体。另外，消除批次效应使得我们能够在不同数据集之间进行查询，从而能够解决单个数据集无法解决的问题。消除批次效应也是提高数据质量，消除数据噪声的有效手段之一[14]。

综上所述，单细胞测序数据中存在的批次效应在 ATAC-seq 数据分析中有着十分重要的影响，批次效应代表了处理不同批次的细胞中出现的技术误差，这种效应可能由测序深度；读长；实验方法或者处理样本的方法，也可能由于取样时间、取样部位、等因素造成，在联合分析不同的实验时，消除批次效应对数据分析的准确性起着重要的作用[14]。

1.4 单细胞数据批次效应消除的复杂性

消除单细胞的批次效应主要被分为两个子任务，“数据整合”（Data integration）和“批次效应消除”（Batch Correction）[17]，这些子任务在需要去除的批次效应复杂性方面存在差异。“批次效应消除”大多数情况下处理同一实验中样本之间的批次效应，其中细胞身份组成是一致的，并且该效应通常是准线性的。相比之下，而“数据整合”处理来自可能使用不同实验生成的数据集之间的复杂且嵌套的批次效应，细胞身份可能在这种情况下的批次之间不共享[18]。

不同的批次效应是否应该被消除取决于实验问题和实验设计，如果我们的数据需要考虑到不同数据的差异，我们则应该在一定程度上保留不同的生物背景导致的批次效应，同样的，当我们想要讨论不同操作方式对数据的影响时，那我们就应该保留由于技术误差导致的批次效应[19]。整合单细胞数据目的是为了将高通量测序的数据集和样本合并，从而生成噪声更低，质量更高的数据以进行下游的后续数据分析，但是根据需要处理的批次效应的复杂度，对数据执行“数据整合”（Data integration）或者是“批次效应消除”（Batch Correction），并选择最我们处理数据所需要的最准确最快捷的方式，来矫正单细胞数据中需要消除的批次效应。

本实验分别设计了两种批次消除任务，分别消除由于不同供体导致的批次效应（3 Samples Task）和由于不同供体、不同取样部位导致的批次效应（4 Samples Task），在处理 3 Samples Task 时，我们期望能够在保留原有的生物特征的前提下最大程度的消除来自不同供体之间导致的差异，以形成质量更高的数据集，而在处理 4 Samples Task 时，由于不同的取样部位可能包含不同的细胞类型，我们则期望能够将不同取样部位导致的细胞间差异保留下来，并消除掉由于不同供体导致的批次效应。

1.5 不同的批次效应矫正模型

消除单细胞数据中的批次效应的方法主要由三个步骤组成：1）降维处理；2）建模并去除批次效应；3）投回到高维空间。在这三个步骤中[14]，通常第二步：建模并去除批次效应是去除批次效应的核心环节，但许多方法主要通过投影到低维空间（第一步）来提高信噪比，并在该低维度下进行批次矫正（Batch Correction）从而得到更好的批次效应消除效果。根据这三个步骤的不同，整合模型被分为多种不同的类别，这些方法可能会利用线性或者非线性降维方法，线性或非线性批次效应模型，输出不同格式的批次矫正数据。

总而言之，我们可以将整合模型大体上分为四类，分别是全局模型（Global Model），线性嵌套模型（Linear embedding Model），基于图像的模型（Graph-based methods）以及深度学习方法（Deep Learning approach）[14]。截止到 2020 年十一月，至少有 49 种用于单细胞测序的消除数据批次效应的方法，但这些方法中的大多数都有着自己的优势和劣势，并不能在所有的情况中都取得比较好的批次效应矫正效果，这些模型中，包括 Harmony，scVI，BBKNN 几种软件在此前的基准测试中对单细胞数据表现出了较好的批次效应消除效果[16,14,20]。

1.5.1 全局模型（Global Model）

全局模型起源于批量（bulk）转录组学，这个模型的核心思想就是经验贝叶斯方法[21]，通过将全局的批次效应这种模型通过将批次效应视作全局存在的一致效应，以 Combat 算法为例：

其核心思想是每个批次的数据都会对数据整体产生一个批次特异的加性和乘性效应，最早被广泛应用于单细胞测序数据批次矫正的一种软件。Combat 可以用以下这个

公式来概括[21]:

$$Y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg}$$

其中，Y_{ijg}代表来自批次i的样品j的基因 g的表达值，α_g是基因的平均表达值，X是样本条件的设计矩阵，β_g是对应于X的回归系数向量，误差项ε_{ijg}服从期望为 0 和方差为σ_g的正态分布$N(0, \sigma_g)$，γ_{ig}和δ_{ig}分别表示批次 i 中基因 g 的加性和乘性批次效应。

这个算法总共分为以下三步：

1）标准化数据：对每个基因，减去其在所有样本中的平均表达量，然后除以其在所有样本中的标准差。

2）估计模型参数：使用经验贝叶斯方法来估计模型中的各参数，包括批次效应和其他可能的协变量效应。

3）调整数据：使用估计的参数对数据进行调整。

作为最早提出的批次效应消除算法，Combat 算法同样为其他算法提供了灵感，但根据之前文章的报道，Combat 算法在计算小数据集中的批次效应时，会引入误差，导致批次效应消除结果不理想[16]，因此，在本文中，我们不讨论全局模型的数据整合结果，主要讨论其他三种模型消除不需要的批次效应的能力。

1.5.2 线性嵌套模型（Linear embedding Model）

线性嵌入模型是一种专门为单细胞测序数据进行数据整合的模型，这种模型主要通过线性方法如奇异值分解（SVD）或者主成分分析（PCA）将单细胞数据将数据嵌入到低维空间中，然后在低维空间中为每个数据寻找其局部邻域，然后通过调整细胞在低维空间中的位置来矫正其批次效应。这种模型的主要原理可以概括为以下几个步骤[14]:

1）通过线性的方法，如奇异值分解（SVD），或者主成分分析（PCA）等方法将数据降维到低维空间中。

2）在低维空间中，根据数据的最邻近值（Nearest Neighbor）跨批次查找相似细胞的邻域然后通过调整数据在低维空间中的位置来矫正批次效应

3）最后，这种模型会将消除批次效应之后的数据投回高维空间中。

常用的线性嵌入模型包括 Harmony 和 MNN，通常情况下，这种计算方法因为使用线性方法将数据降维到低维空间中，对数据的批次效应矫正有着较好的效果。

1.5.3 基于图像的方法（Graph-based methods）

基于图像的方法通常有着最快的运行速度，这种方法使用最邻近图（Nearest Neighbor Graph）的方法来表示数据，这类方法使用最邻近图来表示每个批次的数据，在该图中，每个细胞是一个节点，并强制在细胞之间建立连接，来连接不同的细胞，从而矫正批次效应，在强制连接后，如果两种细胞仍然不相似，这种方法会修剪掉这些强制边来允许细胞类型的差异。BBKNN 就是这类方法中的一个典型例子[22]。

1.5.4 深度学习模型（Deep Learning approach）

近年来，机器学习中的深度神经网络领域经历了巨大的进步。利用这些发展，研究人员开始将神经网络应用于批次效应消除问题，从而产生了移除批次效应的替代方法[23]。这种方法通常都有着较高的准确性，但其在运行时需要占用的计算资源通常远大于其他方法，深度学习方法矫正批次效应的核心原理是利用自动编码器网络，这些方法可以通过条件自分自动编码器在嵌入空间中对维度进行条件降维。总体来说，深度学习方法通过学习数据的复杂特征表示，并结合条件信息或局部线性修正，可以有效地去除批次效应，从而提高数据分析的准确性和可靠性。

1.6 研究目的及意义

批次效应可能会掩盖细胞亚群之间的真实差异。通过消除批次效应，我们可以更准确地识别和比较不同细胞的生物学特征。在大型研究项目中，我们也可以通过消除批次效应将数据整合到一起。本研究旨在比较四种矫正批次效应的算法处理两个不同任务的表现，以评估它们对生物特征的保留和批次效应的移除效果。这两个任务分别由不同供体引起的样本间批次效应以及由不同供体和取样部位引起的批次效应。我们将使用四种批次效应消除算法的不同指标进行比较，以评估这些算法对数据集的生物特征保留和批次效应消除效果。为了确保研究的全面性和可信度，我们还将探讨不同算法在不同任务下的适用性和局限性，为未来类似研究提供参考和启示。

综上所述，本研究将结合领域知识和实际数据分析经验，对比四种批次效应消除

算法在生物学特征保留和批次效应消除方面的优劣，并进一步探讨可能的改进策略和未来研究方向。通过这样的系统评估和分析，我们期望为生物医学领域中单细胞染色质可及性数据中的批次效应处理提供更深入的理解和有效的解决方案，促进相关研究的发展和应用。

2 数据、软件以及实验过程

2.1 使用数据来源

本篇论文主要使用了公开在 GEO 数据库中的来自不同实验室的脑组织单细胞 ATAC 测序数据，如下表表 2.1 所示，这些测序数据来自不同的取样部位或者不同的供体。使用 Sratoolkit 软件将数据对应的.sra 文件下载到服务器上，之后再使用 fasqc 工具将数据转换为.fastq 格式测序文件。

表 2.1 主要实验数据

Table 2.1 Main data

Dataset	GSM Accession	Sample name	Donor	Source
4-Samples task	GSM4441827	Substantia Nigra	donor 1	[24]
	GSM4441821	Caudate	donor 2	
	GSM4441824	Hippocampus	donor 3	
	GSM4441825	Frontal Gyrus	donor 4	
3-Smaples task	GSM7156226	primary motor cortex	donor 5	[25]
	GSM7156218	primary motor cortex	donor 6	
	GSM7156228	primary motor cortex	donor 7	

本研究为了研究不同批次效应消除模型对不同复杂程度的批次效应处理的效果，该研究分别将 7 组数据分为两种类型的任务：4-Sampless task 与 3-Samples task，来分别评估在处理不同复杂度批次效应的过程中，不同的模型的批次效应消除能力如何。

2.2 数据预处理

数据预处理过程使用基于 Nextflow 编写的 pipeline 完成。其主要包括 barcode 的矫正，接头剪切，序列比对，填充文件配对信息，合并 BAM 文件，生成 fragment 文

件等步骤。主要涉及到的程序包括 TrimGalore，BWA-MEM，Samtools 以及 Snapatac2。如使用软件如下表表 2.2 所示：

表 2.2 数据预处理软件

Table 2.2 Softwares for preprocessing

软件名称	版本	Source
NextFlow	21.04.3	[26]
TrimGalore	0.6.10	[27]
BWA-MEM	2.2.1	[28]
Samtools	1.18	[29]
SnapATAC2	2.5.3	[30]
Scib	1.1.4	[16]
Scanpy	1.10.0	[31]
Anndata	0.10.6	[32]

2.2.1 测序接头修剪

TrimGalore 是一个自动-化接头修建和质量控制脚本，可以用于剪去测序文件的接头序列和一些低质量序列，由 Felix Krueger 团队开发，适用于所有的高通量测序分析的数据预处理过程，包括全基因组甲基化测序（WGBS），空间基因组学，染色质可及性分析（ATAC-seq）等数据分析的数据质控流程。该脚本的工作原理主要包括以下三个步骤 1）质量控制：在修剪接头之前，先从 read 的 3’端剪掉低质量的碱基序列；2）接头修剪：Cutadapt 会查找并删除 read 序列的 3’端的接头序列；3）移除短序列：TrimGalore 可以根据 read 的长度过滤掉过短的 reads[27]。

在数据的预处理过程中，首先使用了自己写的脚本对上一步得到的 fastq 文件中的 barcode 进行连接和矫正操作。然后使用 TrimGalore 进行预处理并对部分样本的 barcode 进行剪切，剪切完成后会在指定目录下输出 trimming_report.txt，并生成新的剪切后的 fastq 文件替换之前未剪切测序接头的 fastq 文件。

2.2.2 序列比对

BWA-MEM 是 Burrows-Wheeler Aligner 软件包的一部分，是一个序列比对软件，

可以将低差异序列比对到参考基因组上，BWA-MEM 算法可以用来比对长度范围介于 70bp 到 1Mbp 之间的 read，与其他 BWA 算法相比，BWA-MEM 更快更准确，这也是它目前能被广泛应用的重要原因之一。BWA-MEM 算法基于一种寻找最大精确匹配（maximal exact matches，MEM）的算法。BWA-MEM 算法更有利于用于比对可能包含结构变异或来自具有不同基因组的不同物种的序列的特征。该程序的输出可以直接作为下一步 Samtools 程序的输入文件[28]。

得到上一步的 TrimGalore 剪切处理后得到的 fastqc 文件之后，使用 BWA-MEM 将其与人类参考基因组进行序列比对，将序列匹配到对应的基因组上的的位置，生成对应的 mapping_stats.tsv 文件，以及 bam 文件作为下一步的输出。

2.2.3 Bam 文件合并

除了上面介绍的几种软件，Samtools 软件包在处理和分析高通量测序数据流程中也被广泛使用，Samtools 是一组以 SAM（Sequence Alignment Mapping），BAM 和 CRAM 格式文件进行比对的程序包，由 Heng Li 等人于 2009 年开发，它可以将文件在 SAM，BAM，CRAM 这些格式之间进行转换，排序，合并，以及索引。并且可以快速检索文件中的任何区域并进行读取[29]。

在上一步得到 bam 格式的文件后，这一步使用 samtools 中的 fixmate 命令以及 merge 命令填充比对文件中的配对坐标等信息，并对 bam 文件进行合并处理。最后得到输出的 bam 文件。

2.2.3 生成 fragment 文件以及数据的注释、合并

SnapATAC2 是一款功能强大的单细胞组学数据分析以及处理软件，由张垲等人在 2023 年发布，SnapATAC2 整合并优化了一些常用的数据分析软件，是一款可扩展性十分强大的软件，能够有效处理大规模单细胞组学数据。传统的非线性降维方法通常存在计算效率低、资源占用高的问题，而 SnapATAC2 则通过使用 Rust 编程语言执行计算密集型子程序，并提供 Python 接口，实现了高性能和易用性的完美结合。同时，SnapATAC2 还采用了磁盘数据结构和离线算法，进一步提高了对大规模数据的处理能力，降低了系统资源的负担，使得研究人员可以更轻松地进行大规模单细胞组学数据的分析和挖掘。SnapATAC2 不仅仅提供了一种高效的工具，更重要的是它对单细胞数

据分析的推动作用，它为研究者提供了一款方便快捷的工具有助于他们更好地理解细胞的复杂性，揭示细胞间的关系和调控机制，推动单细胞组学领域的发展和进步[30]。

在得到上一步合并的 bam 文件之后，使用 snapatac2 将 bam 文件转换为 fragment 文件，然后分别根据标志基因的表达量或者 barcode 注释出细胞的类型，然后分别将七组数据合并为两个数据集，储存为 anndata 数据格式进行之后的分析。

2.2.4 数据集的标准化处理

本实验使用 Shifted Logarism 算法来对数据进行标准化处理，这种算法基于 Delta 方法，这种方法使用函数 f(Y)对原始数据 Y 进行处理[33]，从而让整个数据集中的数据之间差异更小，Shifted Logarism 算法通过以下公式来实现：

$$f(y) = log\left(\frac{y}{s} + y_0\right)$$

在这个公式中，y 代表原始数据，s 代表数据集的尺寸系数。y_0则代表伪计数，这一公式中的尺寸系数取决于细胞的不同以及取样方式的不同。通过 Shifted Logarism 算法，我们能够快速的减少数据之间的差异，使得细胞中的批次效应更好被识别出来，便于我们之后实验中的批次效应移除，保证每种批次效应移除算法尽可能准确地计算并移除批次效应。

2.3 数据批次效应矫正

本篇研究中使用的数据批次效应矫正算法主要包括 BBKNN，Harmony，MNN 以及 scVI，这几种方式分别应用了基于图的方法，线性嵌入模型以及深度学习模型的原理，在之前发表的文章中[16,14,20]，这些方法在执行单细胞 RNA-seq 数据的批次效应消除方面有着出色的表现。

2.3.1 Harmony 与 MNN（Linear Embedding Model）

Harmony 与 MNN 方法都属于是线性嵌入模型，Harmony 整合模型的基本算法原理是：该模型首先利用主成分分析（Principal Component Analysis, PCA）算法将每个细胞的测量值降维到二维空间中，然后 Harmony 会接收低维空间中每个细胞的数据然后运行一个迭代算法来调整数据集的批次效应。其中涉及到的迭代算法主要分为以下几

个步骤[34]:

1）首先，Harmony 会利用模糊聚类的方法将每个细胞的测量数据都分配到多个不同的簇中，之所以用模糊聚类，是因为这样的方法可以让数据集之间的聚类差异性最大化。

2）然后，Harmony 为每个聚类簇计算一个全局质心（Global Centroid），同时也计算出每个数据集的全局质心。

3）根据计算得到的质心，Harmony 会再为每个数据集计算出一个矫正因子。

4）最后，Harmony 会根据细胞特异因子（即第一步中通过细胞的聚类结果加权的数据集校正因子的线性组合），来矫正每个细胞的批次效应。

5）Harmony 会一直重复这个过程，直到最终的结果收敛。

而 MNN 也是利用类似这样的原理，不同的地方就是 MNN 在消除批次效应过程中并没有迭代计算过程，而是基于相互最邻近的方法，通过在嵌入空间中寻找不用批次之间的相互最邻近细胞来识别和矫正批次效应[35]。

2.3.3 scVI（Deep Learning approach）

scVI 模型（single-cell variational inference，单细胞变分推断）于 2018 年由 Romain Lopez 等人提出，最先被应用于 RNA-seq 的数据整合中，并且在 RNA 数据的批次效应消除算法的基准测试中表现明显的优势。scVI 作为一种可扩展的框架，可以用于概率表示和分析单细胞中的基因表达数据。但在运行 scVI 算法时，通常需要使用 GPU 进行计算，相较于其他的算法，需要更多的计算资源[14]。scVI 算法进行批次效应矫正主要包括以下步骤:

scVI 首先接受一个包含 n 个细胞和 g 个基因的单细胞测序数据矩阵作为输入，此外，还可以选择提供一个包含 p 个协变量的设计矩阵 S，比如日期，或者供体的信息。然后 scVI 将会使用神经网络来学习一个潜在空间，这个潜在空间可以捕捉到细胞的生物学差异，同时忽略掉不需要的批次效应，最终将消除掉批次效应之后的结果通过解码器返回到高维空间中[36]。

scVI 作为一种深度学习方法，有许多参数需要进行调整来确认最佳的整合办法，在本研究中，我们使用 scVI 中的默认参数完成对该模型的训练。

2.3.4 BBKNN（Graph-based Methods）

BBKNN（Batch Balance K Nearest Neighbors）是一个快速直观的批次消除算法，可以直接通过 Scanpy 进行调用，BBKNN 函数可以在数据内部创建一幅邻居图，用于后续聚类或者 UMAP 可视化分析等操作。

一般的基于图的批次效应消除方法利用 K 最邻（K Nearest Neighbors, KNN）算法为数据结构中的每个细胞找到 K 个最近的邻值，然后将这些邻近值转换为指数相关的 Connectivities，作为进行后续分析的基础。

但如果不同的批次中存在较为严重的人为技术误差， 把不同批次中相关的细胞类型连接起来是非常困难的，而 BBKNN 则通过在每个批次中分别识别 KNN，而不是在整个数据结构中进行识别[22]，例如，当我们设定$K = 3$时，BBKNN 会对第一个批次中的每个细胞识别 3 个最邻近，对于第二个批次，第三个批次，BBKNN 也会做相同的操作。最后，对于每个细胞，BBKNN 会将来自每两个批次之间的最邻近值合并在一起形成一个新的邻近值列表，根据这个列表，我们可以更方便地在两个批次的数据之间建立连接，从而消除批次效应。

2.4 评估指标

正如我们在前文提到的，消除批次效应意味着我们需要分别解决两个问题：生物学差异和技术性批次差异，因此，是否能够解决这两个问题，是评估一组数据批次效应的重要指标，本篇研究中分别使用了图像关联性（Graph Connectivity），k 最邻近批次效应检验（k nearest neighbor Batch Effect Test，kBET）作为评估批次效应是否被移除的指标，并使用调整兰德指数（Adjusted Land Index，ARI），归一化互信息（Normalization Mutual Inofrmation，NMI）作为生物学特征的保留程度的指标。

2.4.1 调整兰德指数

调整兰德指数（ARI）是用于衡量两个聚类结果之间相似性的指标，ARI 通过考虑两个聚类结果中的偶然一致性的可能性来修正兰德指数的随机性敏感性。它的计算过程涉及以下步骤和公式[37]:

假设现在有一个包含 N 个样本的数据集，分别被两种不同的聚类方法 C1 和 C2

进行聚类，要想计算 ARI，首先要计算同时出现在 C1 和 C2 中相同簇的样本对数a，以及 C1 和 C2 中不同簇的样本对数b，然后根据以下公式计算兰德指数：

$$RI = \frac{a + b}{\binom{N}{2}}$$

在这个公式中$\binom{N}{2}$代表N个样本中可能的样本对数。得到兰德指数后，再计算随机聚类的兰德指数的期望值E：

$$E = \frac{\sum_i \binom{n_i}{2} \sum_j \binom{n_j}{2}}{\binom{N}{2}}$$

其中，n_i代表簇i中的样本数，n_j代表簇j中的样本数。最后根据以上结果计算最终的调整兰德指数：

$$ARI = \frac{RI - E}{max(RI) - E}$$

其中$max(RI) = 1$。

根据调整兰德指数定义，我们可以用它来比较两个聚类结果之间的相似性的指标，因为 ARI 会考虑到聚类结果中的偶然性，并给出一个-1 到 1 的值，其中 1 表示两个聚类结果完全一致，0 表示随机聚类效果。在批次效应消除评估中，ARI 可以帮助评估不同批次合并后的数据是否仍然保留了原始细胞之间的区别。如果 ARI 值高，说明不同批次的细胞仍然在新的数据集中保持着相对稳定的聚类模式。

在本研究中，通过调用 scib 包中的 scib.metrics.ari() 函数分别计算了两个数据集消除不需要的批次效应之前 leiden 算法的聚类结果与细胞标签的相似程度，以及消除批次效应之后的聚类结果与细胞标签之间的相似程度。

2.4.2 归一化互信息

归一化互信息（NMI）是对互信息（MI）值的归一化，NMI 可以作为量化聚类质量的一种度量指标。NMI 的结果通常介于 0（无相关性）和 1（完美相关性）之间，NMI 的原理主要依赖于互信息的概念，它可以用来衡量两个变量之间的相互依赖性，在本研究中，这两个变量是指数据集在消除批次效应前后的聚类分配。

NMI 的计算过程主要包括以下步骤[16]：

首先计算每个聚类结果的的熵，熵是衡量聚类结果中的不确定性或混乱程度，对于一种聚类结果C，其熵可以通过以下公式计算：

$$H(C) = -\sum_{i=1}^{k} \frac{n_i}{n} \log \frac{n_i}{n}$$

在这个公式中，n是样本总数，n_i是第i类的样本数，k是类别数，log 通常取 2 或者自然数对数。

然后，计算两个聚类结果之间的互信息，互信息衡量了两个随机变量之间的相关性和相似性，可以用来衡量两个聚类结果之间的相似性，对于两个聚类结果C和C'，他们两个之间的互信息可以用以下公式来表示：

$$I(C, C') = \sum_{i=1}^{k} \sum_{j=1}^{k'} \frac{n_{ij}}{n} log \frac{n \bullet n_{ij}}{n_i \bullet n'_j}$$

最后，可以通过归一化互信息来计算 NMI，NMI 的计算过程如下：

$$NMI(C, C') = \frac{I(C, C')}{\sqrt{H(C) \bullet H(C')}}$$

与 ARI 类似，NMI 也可以用来评估两种聚类结果的相似度，但 NMI 考虑了两个聚类结果中的相对信息量，并给出一个 0 到 1 的值，这个值越接近 1，表示两种聚类结果相似性越高。在评估生物学特征是否被保留时，NMI 可以用来评估批次效应消除前后数据中不同细胞是否能够被聚类算法识别出来。

本研究中，通过使用 scib 中的 scib.metrics.nmi() 函数分别计算了批次效应被消除之前 leiden 聚类结果与细胞标签之间信息的共享程度，以及消除批次效应之后 leiden 聚类结果与细胞标签之间信息的共享程度。

2.4.3 k 最邻近值批次效应检验(kBET)

k 最邻近值批次效应检验（kBET）这个概念最早由 Tran HTN 等人于 2020 年提出[22]，用于评估批次效应，它通过运行一个卡方检验来评估批次效应的概率。计算 kBET 的过程主要包括以下步骤：

首先，对于每个样本我们可以找到其在高维空间的 n 个最近邻，n 是一个参数，在本实验中，我们将其设置为默认值 50，然后，根据我们设定的 k 值，我们需要检查

他们是否来自于同一批次，计算其最近邻来自同一批次的比率，这个比率称为拒绝率，这一拒绝率越低，则说明不同批次中的样本混合的越好，即不同样本间的批次效应被很好的消除了[22]。通过计算与比较不同软件整合之后样本的 kBET 值，我们可以很好的知道样本之间的批次效应是否被很好的消除，从而指出哪种整合方式能够最大程度上的消除批次效应。

本研究中，我们用这一指标评估了两组数据集在进行预处理之后不同批次之间细胞的混合程度以及消除批次效应之后不同批次之间相同细胞的混合程度。

2.4.4 Graph Connectivity

Graph Connectivity（GC）即图连通性，这一指标通过为每一种细胞类型构建一个子图并每个细胞类型标签的子图的连接性。其主要计算过程根据以下公式[38]。：

$$GC = \frac{1}{|C|}\sum_{c\in C}\frac{|LCC(subgraph_c)|}{|c|}$$

公式中 C 代表所有细胞类型的集合，|LCC(subgraph$_c$)|则是细胞类型 c 的子图中最大连接组件的单元数，例如有一个由 10 个细胞组成的子图，其中 7 个细胞可以通过一系列的边相互连接，那么这 7 个细胞就构成了这个子图的“最大连接组件”，而“子图的最大连接组件中的单元数”就是 7。|c|则代表细胞类型 c 的总单元数。

这一公式能够计算得到所有细胞类型平均连接性的平均值，从而为我们评估消除批次效应前后来自相同细胞类型的细胞之间是否有较强的连通性提供了参考，根据这一指标，我们可以用来评估来自不同批次的同类型细胞是否更好的建立了连接。

3 结果

3.1 数据质量

为了保证实验数据的质量，在进行数据的整合和分析之前，有必要对数据进行进一步的预处理，保证数据集的质量，我们首先过滤掉了 tsse 小于 7 的数据，然后对片段对长度分布进行了可视化，最后对两个数据集分别进行了标准化处理。一定程度上保证了最终结果的可信度。

3.1.1 t-SSE 富集得分

我们可以通过 TSS 富集得分与单一片段数量之间的关系来表示过滤后数据的质量，TSS 富集得分这项指标主要用来衡量 ATAC-seq 实验中 Tn5 转座酶的靶向效果，较低的得分代表着 ATAC-seq 的质量较差[39]。从图 3.1 中，我们可以发现在 4-Samples 任务数据集中的四组单细胞 ATAC-seq 数据在过滤后，细胞的片段都富集在转录起始位点附近，说明过滤后的测序数据能够准确反应细胞的基因活性。

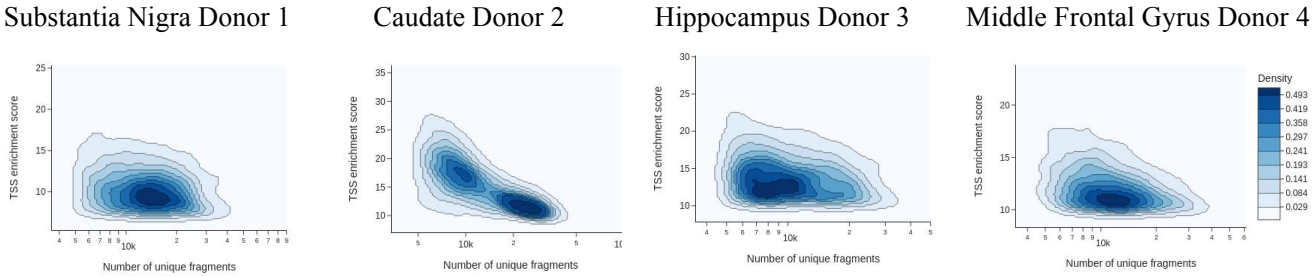


图 3.1 4-Samples 任务数据集过滤后密度图

Figure3.1 filtered 4-Samples task dataset density map

下图 3.2 表示了在 3-Samples 任务数据集的三组单细胞 ATAC-seq 数据在过滤后测序片段数与 TSS 富集得分的密度图，同样的，过滤掉低质量的数据后，细胞的片段都富集在转录起始位点附近，说明过滤后的测序数据能够准确反应细胞的基因活性。

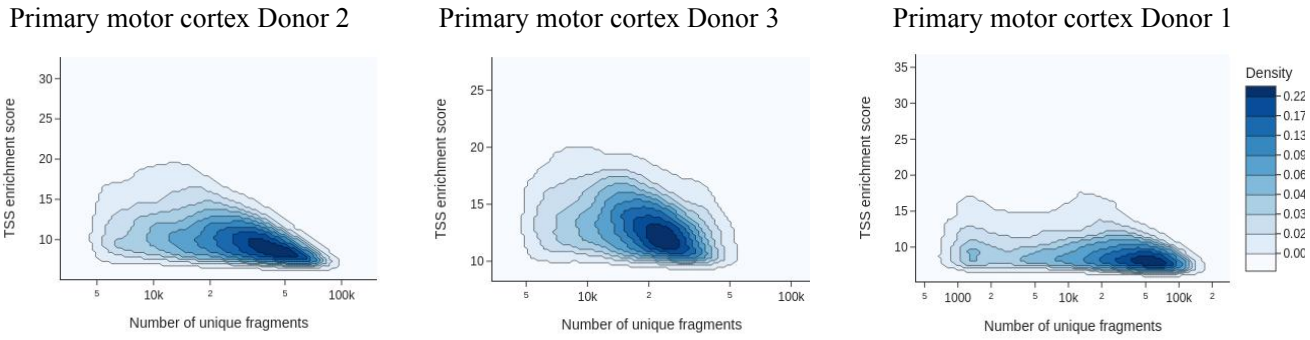


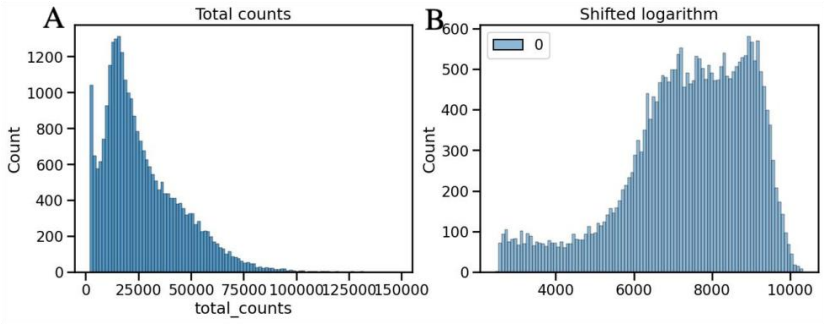
图 3.2 3-Samples 任务数据集过滤后密度图

Figure3.2 filtered 3-Samples task dataset density map

3.1.2 数据标准化

对两组数据集进行质控之后，我们需要将其合并为一个数据集，并对来自不同组的数据进行数据归一化处理，从而降低数据集的方差，有利于后续的降维处理和批次效应矫正过程。

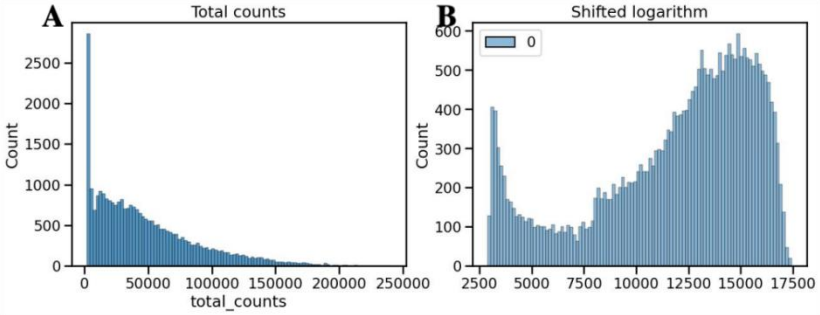
下图 3.3 和 3.4 表示两组数据集进行标准化处理处理前后细胞数与 Reads 数对应关系的变化情况，横坐标为细胞数，纵坐标代表 Reads 数，可以看到在归一化处理之后，数据集内的差异显著降低。



注：*A 图为标准化之前的结果，B 图为标准化之后*

图 3.3 4-Samples 任务数据集标准化前后对比

Figure3.3 Comparison of 4-Samples task dataset before and after standardization



注：A 图为标准化之前的结果，B 图为标准化之后

图 3.4 3-Samples 任务数据集标准化前后对比

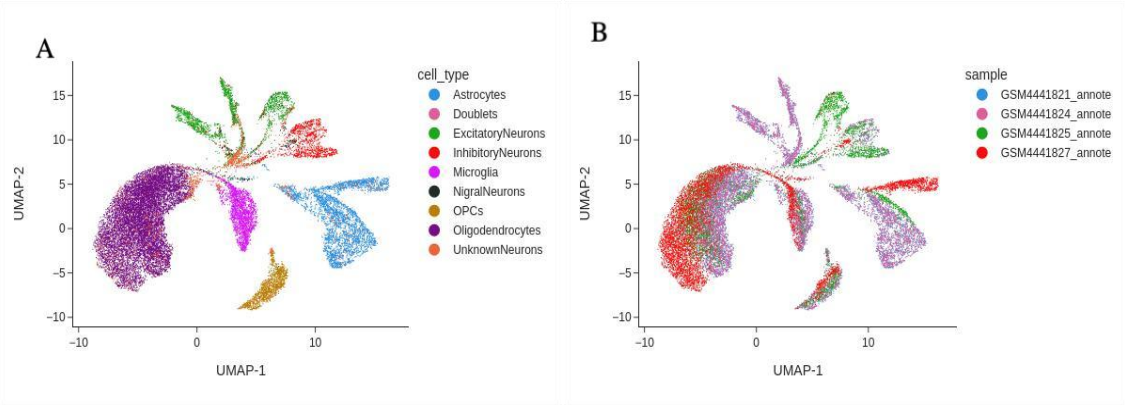
Figure3.4 Comparison of 3-Samples task dataset before and after standardization

3.2 数据整合之前存在的批次效应

3.2.1 未消除批次效应之前来自不同批次的细胞混合效果较差

在对数据进行批次效应消除之前，我们需要知道在合并数据集之前数据集的批次效应程度，我们首先可以通过 leiden 算法根据染色质开放区域来计算细胞间的欧式距离，从而为他们分配聚类标签，然后使用 Umap 降维来对数据进行降维可视化处理。下图 3.5 是对标准化之后，批次效应消除之前的 4-samples 数据集进行 Umap 降维可视化之后得到的结果，可以看到 A 图中来自不同组的细胞类型形成的聚类簇相对松散，细胞之间并没有形成紧密的簇，B 图中也能明显看到来自不同数据集的数据并没有很

好的合并在一起。

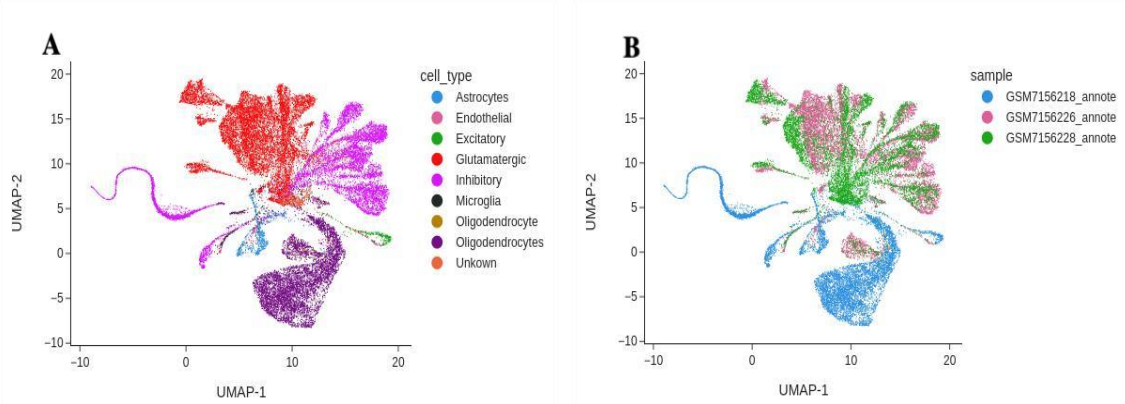


注：A 图用不同的颜色标出了不同的细胞类型，B 图中不同的颜色代表四个数据集，使用数据集对应的 barcode 对细胞类型进行了注释

图 3.5 4-Samples 任务数据集 Umap 可视化

Figure 3.5 Umap Visualization of for 4-Samples Task Dataset

下图 3.6 则表示批次效应消除之前的 3-samples 数据集进行 Umap 降维可视化之后得到的结果，可以观察到来自数据集 GSM7156218 的数据中有很大一部分细胞都是成熟的少突胶质细胞。其他两组数据则包含有其他多种细胞，但是在 Umap 可视化结果中可以看到，这组数据中，相同种类的细胞相对分布的更加紧密，且不同数据集中的数据并没有很好的合并在一起。



注：A 图用不同的颜色标出了不同的细胞类型，B 图中不同的颜色代表四个数据集，使用作者提供的标志基因的表达量确定的细胞的类型，“Olygodendrocyte”与“Olygodendrocytes”分别代表未成熟与成熟的少突胶质细胞。

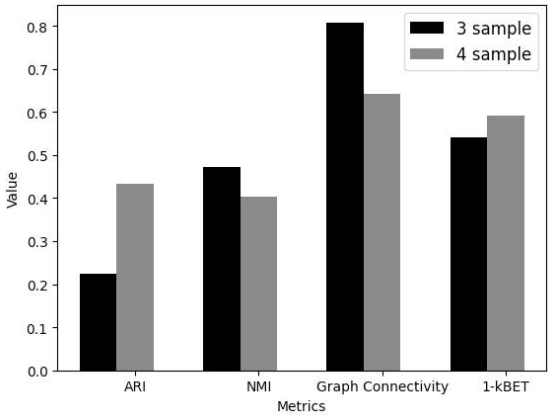
图 3.6 3-Samples 任务数据集 Umap 可视化

Figure 3.6 Umap Visualization of for 3-Samples Task Dataset

3.2.2 未消除批次效应之前两个数据集存在不同的批次效应

在这一部分我们计算了 3-samples 任务数据集和 4-samples 任务数据集两部分数据

的归一化互信息（NMI），调整兰德系数（ARI），图连通性（Graph Connectivity）,kBET 四项指标，并将这四项指标的计算结果进行了可视化，如下图 3.7 所示：



注：此处使用了 $1-kBET$ 作为指标，可以更加直观的看出哪一数据集不同样本间混合的很好

图 3.7 3-Samples 与 4-Samples 任务数据集标准化后各种指标的对比

Figure 3.7 Comparison of Various Indicators after Standardization of 3-Samples and 4-Samples Task Datasets

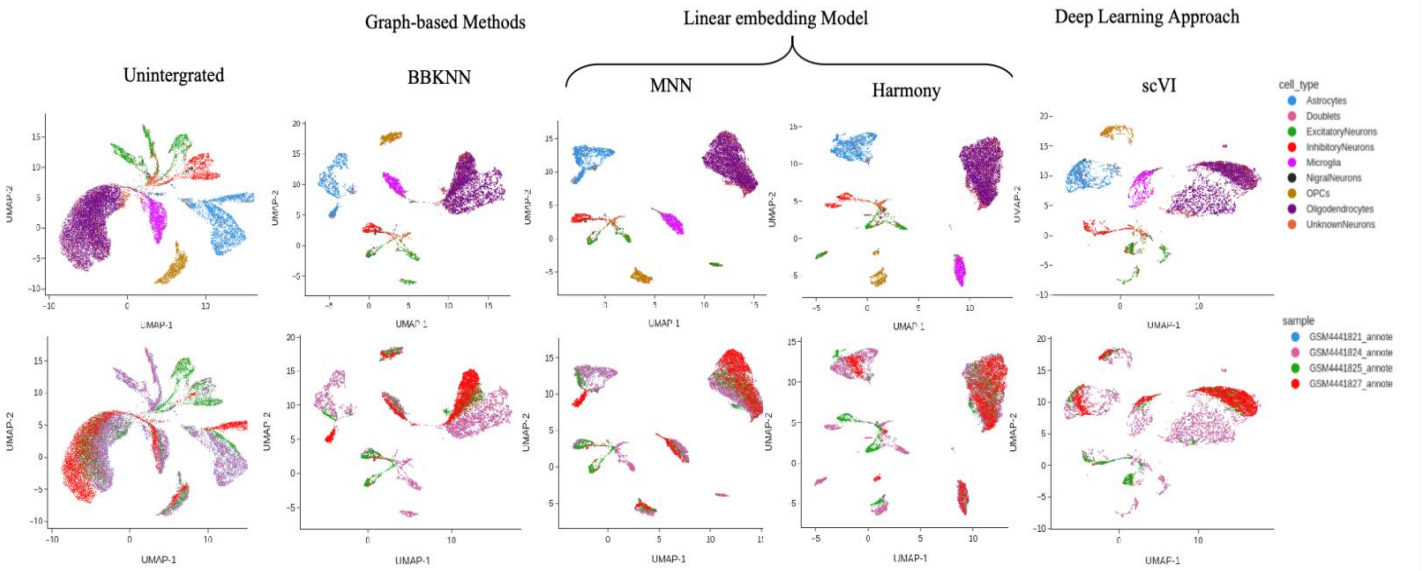
从该图中，我们可以看出，两个数据集的 ARI 和 $1-\text{kBET}$，NMI 和 GC 这些指标都相对较低，这意味着这两组数据集中在未进行对应的批次效应消除操作之前，批次之间的相似程度较高，且生物学特征保留程度相对较低。两个数据集之间各自存在的批次效应并不相同。

3.3 不同方法均明显消除了数据集中的批次效应

分别使用四种不同的方法对数据进行批次效应消除之后对数据集进行 UMAP 可视化处理，并根据细胞类型和实验批次为可视化结果标注颜色。如下图 3.8 和 3.9 所示，上面一行代表不同的细胞类型，下面一行则代表来自不同批次的实验数据。根据所标注的细胞类型，可以看到使用 BBKNN、MNN、Harmony 以及 scVI 等方法之后，相同的细胞类型聚集成了面积更小的簇，且更加集中在了一起。根据不同实验批次标注得到的结果，可以看到混合之后 UMAP 中更多的点重合在了一起。整体上看，两种线性嵌入模型消除批次效应后的可视化后簇的分布比较类似，与另外两种算法表现出了较

大的不同。

下图 3.8 代表 4-Samples 数据集进行批次效应消除之后的结果，可以明显观察到这些高性能的批次效应消除方法都成功消除了不同供体和不同取样部位导致的批次效应，同时保留了细胞类型、染色质开放性等生物学特征。

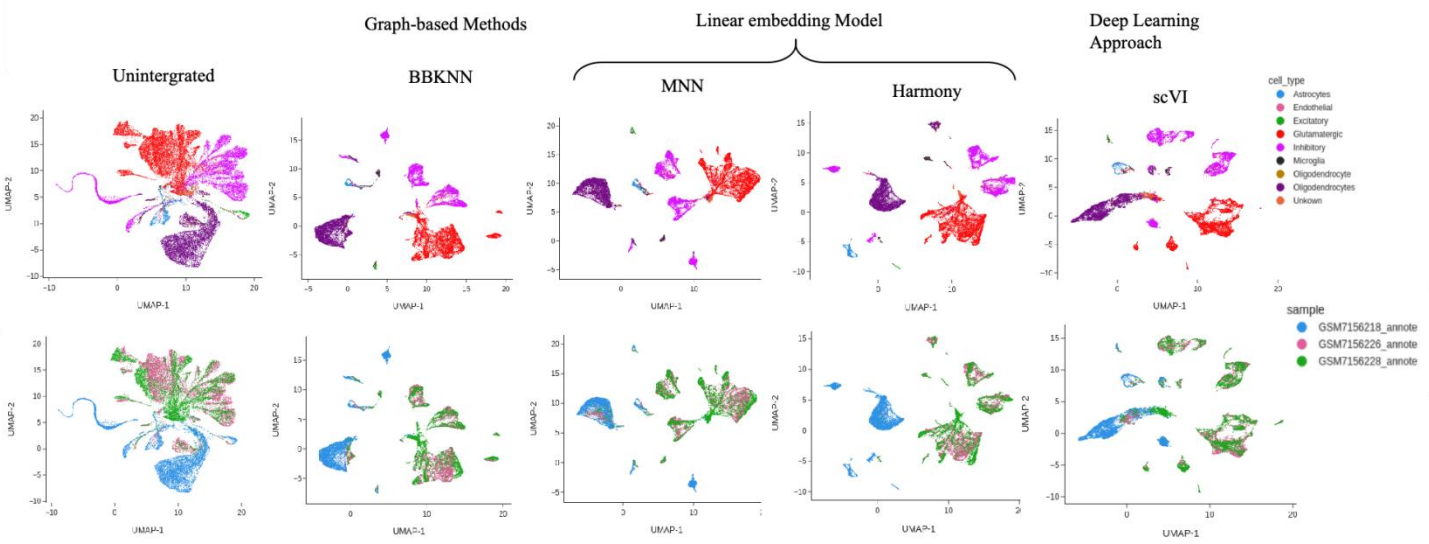


注：两行图分别代表根据细胞类型（上）和批次（下）标注颜色的 Umap 可视化结果图，

图 3.8 4-Samples 任务数据集使用不同软件消除批次效应后的可视化结果

Figure 3.8 Visualization results of using different software to eliminate batch effects on the 4-Samples task

下图 3.9 为 3-Samples 数据集使用不同批次效应消除算法处理后的 Umap 降维图，同样也可以明显观察到这些高性能的批次效应消除方法都成功消除了来自不同供体导致的批次效应，相同类型的细胞聚集的更加紧密，来自不同批次的细胞更好的混合在了一起，同时保留了细胞类型和染色质开放性等生物学特征。在这项任务中，可以看到 GSM7156218 数据集在几种批次效应消除结果中并没有很好的与其他批次数据混合在一起，这是因为在 GSM7156218 数据集中有很大一部分细胞是成熟的少突胶质细胞。



注：两行图分别代表根据细胞类型和批次标注颜色的 Umap 可视化结果图，“Olygodendrocyte”与“Olygodendrocytes”分别代表未成熟与成熟的少突胶质细胞。

图 3.9 3-Samples 任务数据集使用不同软件消除批次效应后的可视化结果

Figure 3.9 Visualization results of using different software to eliminate batch effects on the 3-Samples task

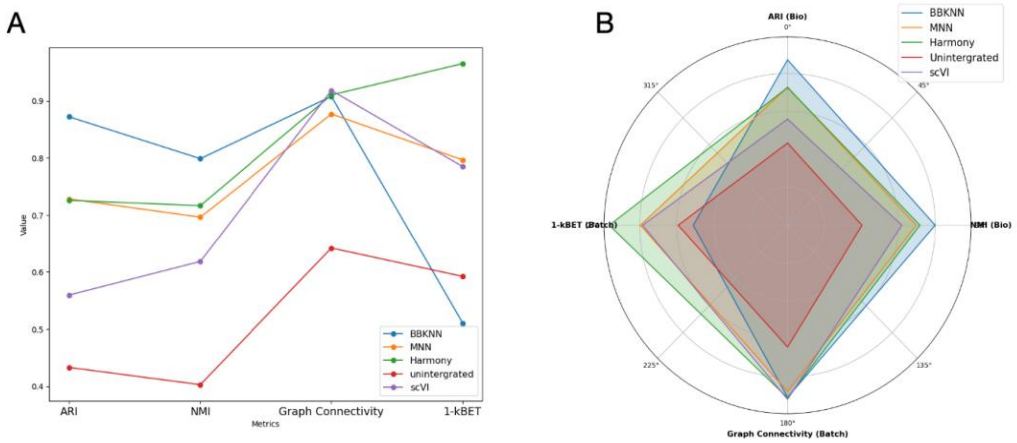
根据以上两个数据集使用不同批次效应消除算法处理后的可视化结果，我们还不能直接判断哪种方法有着最好的表现，要想评估不同方法在处理不同的批次效应任务的能力，还需要更进一步的比较，在本研究中，我们考虑了 ARI 和 1 – kBET，NMI 和 GC 等指标，分别评估四种算法在处理两种批次效应时的批次混合效果与生物特征保留效果。

3.4 Harmony 与 BBKNN 分别在两种任务中有着更好的表现

本研究使用了四种指标来评估不同方法对批次效应的消除效果。如图 3.10 和 3.11 所示，在应用不同批次效应消除算法后，数据集的各项指标明显提升，与未进行批次效应消除的数据相比。我们将这四种指标分为两组：一组用于评估生物学特征（bio-conservation）（如细胞类型）的保留程度，另一组用于评估来自不同批次的细胞是否混合在一起（batch correction）。

图 3.10 展示了对 4-Samples 数据集使用不同批次效应消除方法后的各项指标。可以观察到，BBKNN 算法相较于其他算法更能保留生物学特征，使得细胞类型在消除批次效应后更容易被聚类算法识别。然而，其相对较高的 kBET 值表明，BBKNN 算法并未有效地将来自不同批次的数据混合在一起。Harmony 和 MNN 两种线性嵌入模型算法的多项指标接近，但 Harmony 算法更有效地将不同批次的数据混合在一起。另一方

面，利用深度学习原理的 scVI 算法显著提高了各项指标，表现相对稳定。在 B 图中的雷达图中，我们可以通过计算四种算法各个指标形成的四边形的面积来比较四种算法的综合表现，总体上，BBKNN 在处理不同供体和不同取样部位导致的批次效应时表现较佳，但其混合样本的能力明显低于其他算法。另外三种算法都展现出了较好的批次效应消除能力，并且在消除批次效应的同时也能保留数据中的生物学特征。



注：A 图表示不同方法的各种指标值可视化得到的折线图，B 图为不同方法各种指标的雷达图，因为 kBET 值越低代表越好的批次效应消除效果，此处使用 1-kBET 可以更直观的看出哪一算法有着更好的结果

图 3.10 4-Samples 任务数据集使用不同软件的各种指标可视化结果

Figure 3.10 Visualization results of various indicators using different software on the 4-Samples task

图 3.11 中的折线图表示了四种算法与未处理数据集对比的各项指标表现。可以看出，Harmony 算法在处理 3-Samples 数据集后各项指标明显提高且优于其他算法，而 BBKNN 则展现了与处理 4-Samples 数据集时类似的趋势，前三项指标均有明显提高，但 kBET 的值略微降低，同样表明 BBKNN 算法在将来自不同批次的数据混合在一起方面可能较弱。另外两种算法，MNN 和 scVI，也显著提高了各项指标。根据 B 图的比较，Harmony 和 MNN 在处理由不同供体导致的批次效应中具有明显优势，其次是 scVI 算法，表现稳定；而在这一数据集中，BBKNN 则表现出相对较差的批次效应消除效果。

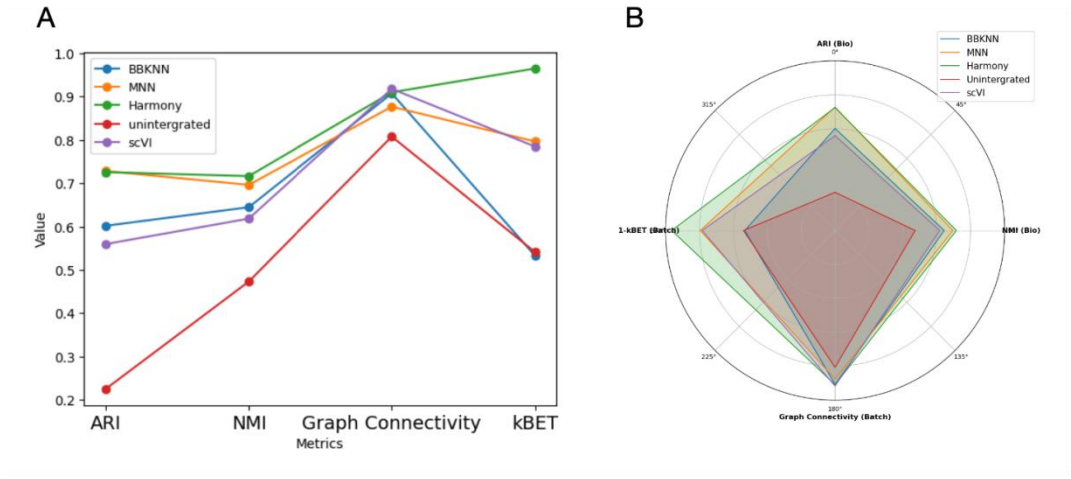


图 3.11 3-Samples 任务数据集使用不同软件的各种指标可视化结果

Figure 3.11 Visualization results of various indicators using different software on the 3-Samples task

4 讨论

在先前的研究中，已经对单细胞 RNA-seq 数据使用不同批次效应矫正算法的效果进行了基准测试。在实验开始前，我们阅读了文献并选择了在单细胞 RNA-seq 数据集上表现优异的四种批次效应消除算法作为我们的基准测试对象[16,14,20]，包括 Harmony、BBKNN、MNN 和 scVI。然而，本文的重点是探讨这些算法在单细胞 ATAC-seq 数据上的批次效应消除效果。处理这种数据时，我们需要考虑到不同的生物学问题，例如如何准确反映出细胞的基因表达情况以及染色质可及性[40]，但目前还没有对这些算法在处理单细胞 ATAC-seq 小数据集中的批次效应进行基准测试的研究。

在本文中，我们主要根据图 4.1 所示的方法进行了对四种批次效应消除算法的基准测试，包括数据准备、批次效应消除和指标比较三个主要部分，并初步得到结果，证明了不同算法的性能高低一定程度上取决于数据集中批次效应的复杂度。有研究表明[16]，Harmony 对于处理更单一来源的批次效应具有更好的适用性。这与我们的结果相符，在我们的研究中，3 个样本数据集存在不同供体导致的批次效应，而 Harmony 在消除这种批次效应时表现良好

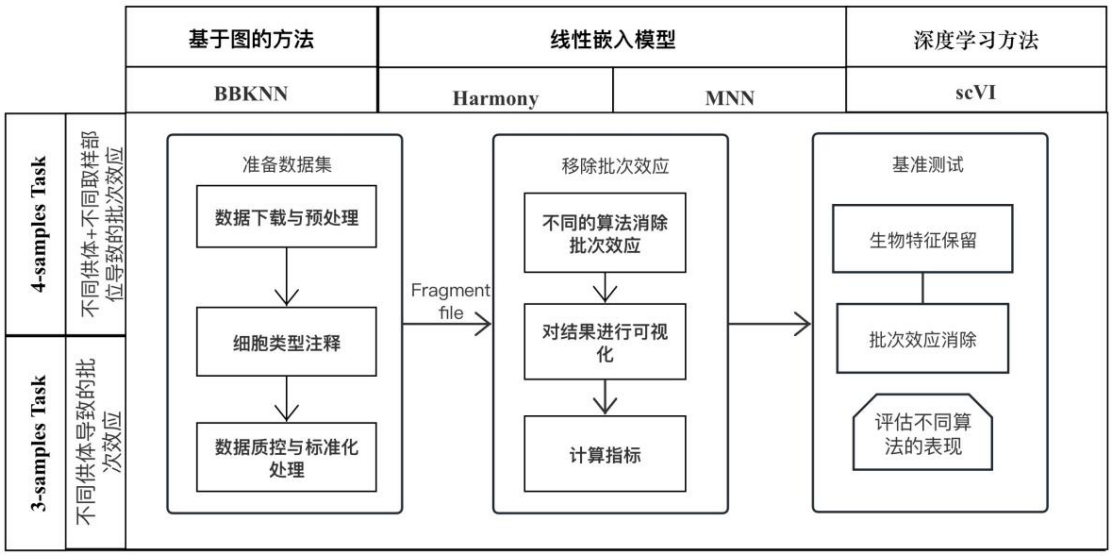


图 4.1 Benchmark 流程图

Figure 4.1 Benchmark flowchart

Harmony 算法基于质心计算每个数据集的校正因子，并使用细胞特定因子对每个细胞进行校正。在数据结构较为简单的情况下，它能够有效识别数据集中的批次效应。然而，在批次效应较为复杂的数据集中，由于数据结构的复杂性，Harmony 可能无法准确找到每组数据的质心，导致在一些指标上不如 BBKNN 算法表现。此外，我们还观察到，虽然深度学习算法 scVI 表现并非最佳，但其相对稳定，通过调整参数可能获得更好的批次效应消除效果，这也展示了深度学习在单细胞数据分析中的潜在应用前景，目前，其他一些深度学习算法如 scGEN，scANVI 等[41,42]，这些方法在消除批次效应中同样有着非常大的潜力。

在进行不同软件的基准测试时，除了考虑算法的效果，计算效率和资源占用也是重要的指标之一。据 Fabian 等人的研究[16]，每种算法的运行速度和内存占用与数据集中细胞和特征的数量相关。不同算法的计算资源和时间消耗差异巨大，因此在选择适合数据分析的方法时，除了批次效应的复杂程度外，还需要综合考虑生物特征保留能力和计算速度，综合考虑这些因素，我们才能选择适用于不同任务的效率最高的计算方法。

结论

本研究使用了公开发布的单细胞不同的 ATAC-seq 数据，经过数据预处理后，将得到的两组 Anndata 数据分别应用了四种不同的批次效应矫正算法。随后对结果进行了分析，发现在处理单一来源批次效应时，Harmony 算法表现较好；然而，在相对复杂的批次效应情况下，BBKNN 算法的多项指标优于 Harmony 算法。BBKNN 显著提高了数据集的生物特征保留指标，但并未有效将样本混合在一起。

此外，深度学习算法在执行本研究中的两项任务时也展现出优异的批次效应移除效果，这证实了深度学习原理在单细胞数据分析中的广泛应用前景。总的来说，本研究通过基准测试验证了四种算法去除批次效应的能力，为单细胞 ATAC-seq 数据分析过程中工具的选择提供了参考，并为单细胞数据处理流程提供了新的思路。

致谢

在这篇论文的最后，我要衷心感谢所有在我学术道路上给予我支持和帮助的人们。你们的付出和帮助使得我能够顺利完成这篇论文，同时也在我学术和人生的道路上留下了深刻的印记。

首先，我要特别感谢我的毕设校外指导老师张垲老师。他不仅为我提供了在西湖大学学习和完成毕业论文的机会，还以其丰富的生物信息学相关经验和专业知识，给予了我许多宝贵的建议和指导。他对于我研究课题的深入理解和耐心指导，使我能够更加全面地掌握研究方法和技巧，为论文的顺利完成奠定了坚实的基础。其次，我也要感谢郝慧芳老师。在我本科阶段，她给予了我非常多的鼓励和支持，给了我学习和成长的机会，激发了我对于生物学研究的热情和动力。

特别要感谢来自 Zhang Lab 的樊佳琪师姐和李泊成师兄。他们在我毕业设计的过程中提供了巨大的帮助，不仅耐心解答了许多复杂问题，还提出了宝贵的建议和意见。他们对于研究方法和技术的熟练掌握和分享，为我研究工作的顺利进行提供了重要保障。同时，也要感谢来自丁一航师兄和刘悦师姐。他们在研究过程中给予了我许多启发和指导，帮助我更加清晰地思考和分析研究问题。他们的经验和见解，对于我论文的改进和完善起到了关键作用。也为我的研究工作提供了丰富的资源和支持。他们的热情和专业精神，激励着我不断追求卓越，不断挑战自我。此外，我也要特别感谢我的女朋友刘思敏。在整个毕业设计的过程中，是她给予了我支持与鼓励，让我相信自己可以克服一切困难。

最后，我要感谢我的朋友与家人。他们在我学术与生活上遇到困难时给予了我无私的支持和鼓励，让我坚定地走在学术道路上。他们的友情和支持，是我前进的动力和勇气的源泉。

再次衷心感谢你们 在我学术道路上的陪伴与支持，使我能够顺利完成这篇论文。您们的帮助和鼓励，是我宝贵的财富和动力，我会倍加珍惜这段宝贵的经历和你们的友谊。愿我们在未来能够继续携手共进，共同探索生命科学的精彩之处。

参考文献

[1] Goodwin, S., McPherson, J. & McCombie, W. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17, 333–351 (2016).

[2] Bell, O., Tiwari, V., Thomä, N. et al. Determinants and dynamics of genome accessibility. Nat Rev Genet 12, 554–564 (2011).

[3] Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145-66.

[4] Kouzarides T. Chromatin modifications and their function. Cell. 2007 Feb 23;128(4):693-705.

[5] Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc. 2010 Feb;2010(2): pdb.prot5384.

[6] Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr Protoc Mol Biol. 2015 Jan 5;109:21.29.1-21.29.9.

[7] Grandi FC, Modi H, Kampman L, Corces MR. Chromatin accessibility profiling by ATAC-seq. Nat Protoc. 2022 Jun;17(6):1518-1552.

[8] Li YE, Preissl S, Miller M, et al. A comparative atlas of single-cell chromatin accessibility in the human brain. Science. 2023;382(6667):eadf7044.

[9] Li, H., Sun, Y., Hong, H. et al. Inferring transcription factor regulatory networks from single-cell ATAC-seq data based on graph neural networks. Nat Mach Intell 4, 389–400 (2022).

[10] Satpathy AT, Granja JM, Yost KE, et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. Nat Biotechnol. 2019;37(8):925-936.

[11] Baldi S. Nucleosome positioning and spacing: from genome-wide maps to single arrays. Essays Biochem. 2019;63(1):5-14.

[12] Virshup I, Bredikhin D, Heumos L, et al. The scverse project provides a computational ecosystem for single-cell omics data analysis. Nat Biotechnol. 2023;41(5):604-606.

[13] Hiltemann S, Rasche H, Gladman S, et al. Galaxy Training: A powerful framework for teaching!. PLoS Comput Biol. 2023;19(1):e1010752.

[14] Heumos L, Schaar AC, Lance C, et al. Best practices for single-cell analysis across modalities. Nat Rev Genet. 2023;24(8):550-572.

[15] van den Brink SC, Sage F, Vértesy Á, et al. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. Nat Methods. 2017;14(10):935-936.

[16] Luecken MD, Büttner M, Chaichoompu K, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41-50.

[17] Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15(6):e8746.

[18] Zappia L, Phipson B, Oshlack A. Exploring the single-cell RNA-seq analysis landscape with the

scRNA-tools database. PLoS Comput Biol. 2018;14(6):e1006245.

[19] Korsunsky, I., Millard, N., Fan, J. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods 16, 1289–1296 (2019).

[20] Swamy VS, Fufa TD, Hufnagel RB, McGaughey DM. Building the mega single-cell transcriptome ocular meta-atlas. Gigascience. 2021;10(10):giab061.

[21] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-127.

[22] Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020;36(3):964-965.

[23] Tran HTN, Ang KS, Chevrier M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21(1):12.

[24] Corces MR, Shcherbina A, Kundu S, et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. Nat Genet. 2020;52(11):1158-1168.

[25] Zemke NR, Armand EJ, Wang W, et al. Conserved and divergent gene regulatory programs of the mammalian neocortex [published correction appears in Nature. 2024 Jan;625(7996):E26]. Nature. 2023;624(7991):390-402.

[26] Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316-319.

[27] Felix Krueger, et al. Felixkrueger/trimgalore: V0.6.10 - Add Default Decompression Path. 0.6.10, Zenodo, 2 Feb. 2023,

[28] Vasimuddin Md, Sanchit Misra, Heng Li, Srinivas Aluru. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. IEEE Parallel and Distributed Processing Symposium (IPDPS), 2019

[29] Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. Twelve years of SAMtools and BCFtools. Gigascience. 2021 Feb 16;10(2):giab008.

[30] Zhang K, Zemke NR, Armand EJ, Ren B. A fast, scalable and versatile tool for analysis of single-cell omics data. Nat Methods. 2024;21(2):217-227.

[31] F. Alexander Wolf, Philipp Angerer, Fabian J. Theis Genome Biology 2018 Feb 06.

[32] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, F. Alexander Wolf bioRxiv 2021 Dec 19.

[33] Dorfman, R. A. (1938). A note on the !d-method for finding variance formulae. Biometric Bulletin, 1, 129–138.

[34] Korsunsky I, Millard N, Fan J, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16(12):1289-1296.

[35] Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421-427.

[36] Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053-1058.

[37] William M. Rand (1971) Objective Criteria for the Evaluation of Clustering Methods, Journal of the American Statistical Association, 66:336, 846-850

[38] Khuller, S., Raghavachari, B. (2016). Graph Connectivity. In: Kao, MY. (eds) Encyclopedia of Algorithms. Springer, New York, NY.

[39] Kazachenka A, Bertozzi TM, Sjoberg-Herrera MK, et al. Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. Cell. 2018;175(5):1259-1271.e13.

[40] Liu J, Ma J, Wen J, Zhou X. A Cell Cycle-aware Network for Data Integration and Label Transferring of Single-cell RNA-seq and ATAC-seq. Preprint. bioRxiv. 2024;2024.01.31.578213.

[41] Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019;16(8):715-721.

[42] Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol. 2021;17(1):e9620.