Biophysics 生物物理学, 2023, 11(2), 30-43
Published Online May 2023 in Hans. https://www.hanspub.org/journal/biphy
https://doi.org/10.12677/biphy.2023.112003

Hans汉斯

单细胞数据库建设的研究进展

陈玲玲[1]，程　烽[1]，胡　桓[1,2]，徐　飞[1,2]，李　翔[1]，林　海[2]

[1]厦门大学物理科学与技术学院，福建　厦门
[2]中国科学院大学温州研究院，浙江　温州

收稿日期：2023年3月29日；录用日期：2023年5月15日；发布日期：2023年5月22日

摘　要

近年来，随着以单细胞转录组测序(Single-cell RNA sequencing, scRNA-seq)技术为重点的大规模生物学实验的兴起，研究人员可以在细胞水平上展开更加深入的研究。基于scRNA-seq技术的优势，尤其是其对研究细胞异质性的能力，越来越多的单细胞数据库涌现出来，为疾病的发生和治疗提供了研究基础，特别是对于复杂的癌症和当前难以完全解决的COVID-19问题。随着scRNA-seq技术的不断发展，单细胞数据库也在不断完善和扩大，涵盖越来越多的物种数据信息，同时提供多种分析功能，为单细胞研究提供了便利。本文回顾了目前广泛使用的单细胞数据库，并对其数据量和数据类型等做了概括总结。此外，我们还调查了研究人员在数据分析方面的使用情况，并得出了单细胞数据库建设的最新进展。最后，本文还针对目前单细胞数据库存在的局限性提出了一些改进建议。

关键词

scRNA-seq，数据库，单细胞分析，标记基因，COVID-19

Research Progress on Single-Cell Database Construction

Lingling Chen[1], Feng Cheng[1], Huan Hu[1,2], Fei Xu[1,2], Xiang Li[1], Hai Lin[2]

[1]College of Physical Science and Technology, Xiamen University, Xiamen Fujian
[2]Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou Zhejiang

Received: Mar. 29th, 2023; accepted: May 15th, 2023; published: May 22nd, 2023

Abstract

In recent years, with the rise of large-scale biological experiments that focus on single-cell RNA sequencing (scRNA-seq) technology, researchers can conduct more in-depth studies at the cellular

文章引用: 陈玲玲, 程烽, 胡桓, 徐飞, 李翔, 林海. 单细胞数据库建设的研究进展[J]. 生物物理学, 2023, 11(2): 30-43.
DOI: 10.12677/biphy.2023.112003

level. Based on the advantages of scRNA-seq technology, particularly its ability to study cell heterogeneity, an increasing number of single-cell databases have emerged, providing a research foundation for the occurrence and treatment of diseases, especially for complex cancers and the currently unsolved COVID-19 problem. As scRNA-seq technology continues to develop, single-cell databases are also constantly improving and expanding, covering more and more species data information, while providing multiple analysis functions, facilitating single-cell research. This article reviews currently widely used single-cell databases and summarizes their data volume and data types. In addition, we investigated the usage of researchers in data analysis and obtained the latest progress in the construction of single-cell databases. Finally, this article proposes some improvement suggestions for the limitations of current single-cell databases.

Keywords

Single-Cell RNA Sequencing, Database, Single-Cell Analysis, Marker Gene, COVID-19

1. 引言

自从 2009 年单细胞转录组测序 scRNA-seq (Single-cell RNA sequencing)技术被首次报道以来[1]，该技术在过去十几年中得到了快速发展。scRNA-seq 技术通过逆转录、扩增和测序等步骤，获得单个细胞内所有 RNA 的表达谱，实现了对单细胞的转录组测序。这一技术为破译细胞不同功能状态提供了前所未有的机会。通过研究单细胞转录组的异质性[2] [3]，scRNA-seq 技术为发育生物学、神经科学、肿瘤学和免疫学等领域提供了许多基本问题的解答[4] [5]。同时，研究单细胞转录组的异质性也是研究肿瘤微环境、异质性、发病机制、转移以及多种肿瘤的侵袭和诊断的重要基础[6]，特别是在研究 COVID-19 方面，scRNA-seq 技术具有不可或缺的作用[7]。此外，许多研究人员使用 scRNA-seq 技术鉴定疾病中基因表达的特异性。例如，在阿尔茨海默病中，研究人员通过该技术鉴定出多个神经元细胞亚群的差异表达基因[8]，在慢性粒细胞白血病中，研究人员表征不同阶段癌症干细胞亚群的分子特征[9]。该技术还揭示了 2 型糖尿病中细胞类型特异性表达的变化[10]。scRNA-seq 技术使得研究人员能够从新的角度看待发育、肿瘤等问题[11] [12]。

随着 scRNA-seq 技术的发展，单细胞数据量呈爆发式增长。目前已经出现各式各样的数据库，覆盖不同物种、组织、细胞类型和健康状态等。对于长期困扰人类的癌症问题，癌症数据库成为科学家们的热门关注对象。癌症单细胞数据库提供丰富的基因表达数据和元数据，为癌症分子机制研究、靶向治疗研发和个性化医疗应用等领域提供有力支持。例如，CCLE (Cancer Cell Line Encyclopedia)数据库提供了单细胞水平的转录组数据[13]，可以用于揭示癌症细胞系的异质性和复杂性，从而帮助开发更有效的治疗方法。癌症单细胞数据库也促进了跨学科合作，推动了癌症研究的发展。另外，自从 2019 年末 COVID-19 在全球爆发以来，单细胞数据库也成为了研究 COVID-19 的重要资源[14]，提供了大量基因表达数据和元数据。单细胞数据库还有助于了解 COVID-19 感染免疫细胞的类型、分化状态和表型特征[15]。此外，单细胞数据库还可以为药物开发和个性化治疗提供重要参考，有助于缓解 COVID-19 的传播和控制疫情[16]。单细胞数据库的应用也有助于深入了解不同类型细胞在感染过程中的变化。例如，单细胞数据库可以帮助研究人员确定病毒感染过程中 T 细胞的活化和亚群分布，以及不同类型细胞中的抗病毒反应等[17]。

这些信息将有助于研究人员进一步了解 COVID-19 感染的机制和病毒与人体细胞之间的相互作用，为治疗和预防 COVID-19 感染提供有价值的信息。随着单细胞数据分析技术的不断发展，科学家已经开始注重单细胞数据分析结果的利用[18]，许多单细胞数据库也提供了分析功能，有利于加强使用者对于数据的理解。同时，部分数据库还提供在线分析功能，让初学者也能够快速获取分析结果。

尽管 scRNA-seq 技术已经得到充分发展，但目前存在的单细胞数据库仍然不够全面，数据的收集及有效利用仍然是一个挑战。随着人们对分析技术的要求不断提高，现存数据库的分析功能也难以达到人们的要求。本文将对目前广大研究者常用的单细胞数据库及单细胞数据分析方法进行介绍。

2. 单细胞数据库质量的评价标准

随着 scRNA-seq 技术的广泛应用，scRNA-seq 数据库已经得到大量积累，但目前鲜少有人对单细胞数据库的优劣做出评价。评价单细胞数据库的优劣应主要考虑以下几个方面：

1) 数据量和覆盖度。数据量和覆盖度是评估单细胞测序数据库质量的重要指标。数据量指的是数据库中可供检索的单细胞测序数据的总量，而覆盖度指的是数据集覆盖的细胞类型、组织类型、物种等方面的广泛程度。这两个指标可以共同反映数据库的数据资源丰富程度和实用性。

2) 数据库的分析功能。数据库的分析功能也是评估单细胞测序数据库质量的重要因素。数据库应该提供高效、准确和可重复的分析方法，以帮助研究人员从复杂的单细胞数据中提取有用的信息。

3) 数据更新频率及数据库的数据格式等。单细胞技术正在不断发展，新的单细胞数据集不断涌现，数据库的更新频率应该高，以便研究人员能够及时获得最新的数据。同时单细胞数据有多种格式，如 10x Genomics、Drop-seq 等，数据库应该提供多种格式的数据，以便研究人员可以使用他们自己的工具和分析流程进行数据分析。

总的来说，一个高质量的单细胞数据库应该具备覆盖度广的大规模的数据集、易于使用的分析工具、高更新频率、可靠的数据来源信息和多种数据格式。

3. 单细胞数据库的数据量统计

随着单细胞技术的发展，单细胞测序数据的规模和数量已经大幅度增加，这也促进了单细胞数据库的发展和壮大。单细胞测序数据库中的数据量越大、覆盖度越广就能够更全面地了解细胞的多样性、功能和相互作用。这对于理解生物学和研究疾病至关重要。因此，有必要调查单细胞测序数据库的数据量和覆盖度情况。

单细胞数据库包含多种类型。例如，SCovid19 数据库是一个针对新冠病毒(COVID-19)感染的 scRNA-seq 数据库[19]。该数据库记录了感染 COVID-19 的 10 个人体组织的 21 个单细胞数据集的 1,042,227 个单细胞。CancerSEA 数据库是一个针对肿瘤单细胞转录组的数据库[20]，结合了 49 个癌症相关的 scRNA-seq 数据集的 41,900 个单细胞，涵盖了包括肝癌、结肠癌、乳腺癌、肺癌等多种癌症类型。HCA (Human Cell Atla)数据库是一个多维度的、开放访问的数据库[21]，它包含了成千上万个人类细胞的转录组、表观组和蛋白质组数据[22]，以及细胞图像和元数据。目前为止 HCA 数据库已经收录人类 80 个组织超过 1990 万个细胞。相比于其他数据库，HCA 数据库具有丰富的细胞类型覆盖范围和高质量的数据，其涵盖了包括不同器官、组织、发育阶段和疾病状态下的细胞类型，可以提供更加全面和完整的人类细胞图谱，并且其使用最先进的 scRNA-seq 技术和标准化的实验和数据处理流程，以保证数据的高质量和可比性。CellMarker 数据库是一个用于细胞类型标记基因鉴定的在线数据库[23]。CellMarker2.0 是 CellMarker 的更新版本[24]，其标记基因数据来源于多个公开的单细胞 RNA 测序数据集，提供人类和老鼠不同组织中不同细胞类型的标志物。其中包含了 355 种癌症类型相关的细胞标记，涵盖了人类和小

鼠的超过 150 种细胞类型和亚型的标记基因信息，包括免疫细胞、神经细胞、心肌细胞等，能够满足不同细胞类型的标记基因查询需求。这些数据库对单细胞的研究做出了巨大贡献。目前常用数据库对应数据信息如表 1 所示。

Table 1. Database data volume
表 1. 数据库数据量

数据库	类型	物种	数据集量(个)	组织种类(种)	细胞量(个)	细胞种类(种)
SCovid	COVID-19 疾病	人类	21	10	1,042,227	/
CancerSEA	25 种癌症	人类	49	/	41,900	/
SC2disease	25 种疾病	人类	/	29	/	341
CellMarker2.0	标记基因	人类、小鼠	/	656	/	2578
Expression Atlas [25]	综合	人类、小鼠等 18 个物种	/	/	590 万	/
HCA	综合	人类	200	80	1990 万	200
CancerSCEM [26]	20 种人类癌症	人类	208	/	/	/
Cellatlas [27]	综合	小鼠等 15 个物种	/	/	260 万	/
HTCA [28]	综合	人类	3000	/	230 万	/
HUSCH [29]	综合	人类	185	45	300 万	270
ABC portal [30]	血液/免疫系统	人类、小鼠	198	/	/	/
MCA [31]	综合	小鼠	/	56	/	160
Tabula Muris [32]	综合	小鼠	/	20	10 万	/
Allen Cell Types Database [33]		人类、小鼠	/	/	27,000	60
The Human BioMolecular Atlas Program (HuBMAP) [34]	综合	人类	1480	31	/	/
Tabula Muris Senis [35]	综合	老年小鼠	/	21	151,882	60
PanglaoDB	综合	人类、小鼠	/	258	559 万	/
scRNASeqDB [36]	综合	人类	200	/	/	8,910
CellBlast	综合	人类、小鼠、斑马鱼等	168	/	/	/

其中，“/”代表原数据库未统计的内容或者本文未收集到相关内容。

由上表可以看出目前单细胞数据库大多侧重点不同且单个数据库数据不够全面，例如，CancerSEA 只收集人类癌症的信息；SCovid 数据库仅限于新冠病毒感染下的单细胞转录组数据，不适用于其他疾病或正常细胞的单细胞 RNA 测序数据，其数据量相对较少，它不能完全覆盖病毒感染下的所有细胞类型和状态；CellMarker2.0 专门研究人类和小鼠的标记基因，但并不是所有细胞类型都有明确的标记基因。该数据库存在部分细胞类型的标记基因缺失的情况。虽然随着单细胞数据急剧增加，非洲爪蟾、斑马鱼胚胎以及秀丽隐杆线虫等细胞数据已经进入我们的视野，丰富了我们对不同物种细胞层次结构的认识，但这些物种的数据资料稀缺，这对相关单细胞研究造成了阻碍。此外，细胞标记基因对细胞注释意义重大，例如 PanglaoDB 和 CancerSEA 的数据库已经从可用的文献信息中获取不同细胞类型的基因用于细胞簇的

注释[37]。然而，这些数据库中的标记信息具有一定的局限性，标记的组织来源、类型和测序技术不足，无法提高细胞注释的准确性。CellMarker2.0 也缺少对除人类和小鼠以外物种标记基因的收集。单细胞数据是单细胞研究的基础，相关数据的稀缺及片面性为单细胞研究的突破增加了难度。因此，未来数据库的建设仍要考虑继续扩大数据量的问题，不仅扩大多物种的多种健康状态的单细胞数据，还应扩大多物种的标记基因信息。

4. 单细胞数据库分析方法及分析工具统计

4.1. 单细胞数据分析流程及分析方法

单细胞转录组的数据分析主要分成预处理和下游分析，分析流程如图 1 所示。预处理过程中原始测序数据经过处理得到分子计数矩阵，计数矩阵中的每个数值代表细胞中一种 mRNA 分子被成功捕获、逆转录和测序的数量[38]，质量控制用于保证下游分析时数据质量足够好，标准化是对细胞计数数据进行缩放等以获得细胞之间可比的相对基因表达丰度[39]，去除实验过程中随机性的影响。数据校正的目的就是进一步去除技术因素和非关注的生物学混杂因素，例如批次、dropout 或细胞周期不同带来的影响[40]。

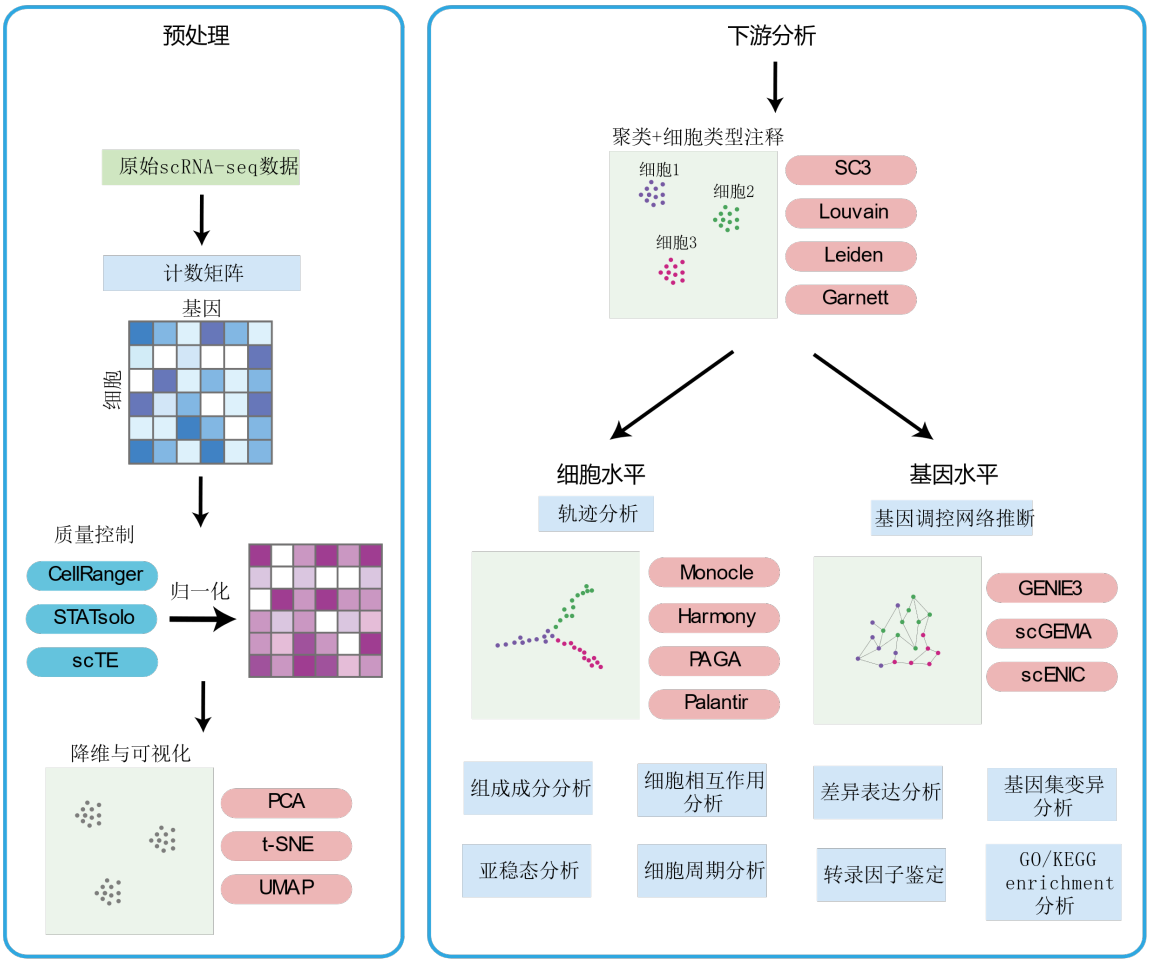

Figure 1. Classic workflow for single cell analysis
图 1. 单细胞分析经典流程

为了减轻下游分析工具的计算负担、减少数据中的噪声并方便数据可视化，预处理过程中通常使用

多种方法来对数据集进行降维。降维的第一步通常是特征选择，即对数据集基因进行过滤保留对数据的变异性具有信息贡献的基因。特征选择后，可以通过专用的降维算法进一步对单细胞表达矩阵进行降维。常用的降维方法包括主成分分析 PCA (Principal component analysis)和 diffusion maps 等。scRNA-seq 数据可视化的最常见的降维方法是 t-SNE (t-distributed stochastic neighbour embedding)和 UMAP (Uniform Manifold Approximation and Projection) [41]。

经过预处理后，下游分析的方法指应用于生物学发现并描述潜在的生物学系统的方法，可分为细胞水平和基因水平的方法。聚类分析是将细胞聚类成簇，通常是任何单细胞分析的第一个中间结果，使我们可以推断成员细胞的身份。聚类方法主要包括 K-means，Hierarchical clustering，Density-based clustering 和 Graph clustering 等[42] [43] [44] [45]。细胞簇在细胞层面可以根据其组成结构分析聚类数据，围绕不同样品落入每个细胞簇的细胞比例进行分析。轨迹推断用于捕获细胞身份之间的过渡状态、不同的分化分支或生物学功能的渐进式非同步变化[46]。在整个伪时间范围内连续平稳变化的基因是决定轨迹的关键基因，可用于识别潜在的生物学过程。细胞间相互作用分析用于研究单个细胞之间的相互作用和细胞间通信的机制[47]，该方法可以帮助研究人员更深入地了解细胞间的交流方式和调节机制，以及不同细胞类型和状态之间的相互作用。由于 scRNA-seq 数据具有高度异质性和噪声性，因此在进行细胞间相互作用分析时，应特别关注数据的质量和可靠性，并采用多重比较校正方法进行结果的纠正和验证。此外，还应结合其他生物学信息，例如 GO/KEGG enrichment 分析和转录因子调控分析等[48]，进一步探究细胞间相互作用的生物学意义。scRNA-seq 数据的周期分析是一种用于确定单个细胞在不同细胞周期阶段的方法[49]，细胞周期是细胞从分裂到下一次分裂的一系列过程，包括 G1、S、G2 和 M 四个阶段。在 scRNA-seq 数据中，通过检测细胞周期相关基因的表达，可以确定单个细胞所处的细胞周期阶段，从而更好地理解细胞状态和细胞功能。除了上述方法之外，在细胞水平上的分析方法还包括细胞组成分析、基因表达量的动态变化以及细胞亚稳态分析等。

基因层面的分析会提供更多的信息，如差异表达分析和基因调控网络推断，不是表面上研究细胞异质性，而是基于异质性探索基因表达相关的原因。scRNA-seq 数据的差异表达分析是指通过对不同条件下的 scRNA-seq 数据进行比较，识别在不同条件下表达差异显著的基因。差异表达分析可以通过比较不同发育时期或不同组织中的 scRNA-seq 数据，发现在不同发育阶段或不同组织中表达差异显著的基因。差异表达分析可以揭示细胞的分化状态和发育轨迹并发现调控细胞功能的关键基因，还能为转录因子分析及细胞间相互作用分析等提供基础数据。基因集变异分析是另一种基于表达谱的分析方法，可用于评估某个基因集在样本中的相对富集程度。在 scRNA-seq 中，基因集变异分析可以帮助确定不同细胞类型之间的基因表达差异以及不同通路的活性水平。GO/KEGG enrichment 分析用于寻找在单细胞水平上特定基因集合的富集情况。在进行 GO/KEGG enrichment 分析时，需要将感兴趣的基因集合与 GO/KEGG 数据库中的注释信息进行比对，然后计算基因集合在每个功能分类或通路中的富集程度。常用的计算方法包括 Fisher’s Exact Test、Hypergeometric Test 和 GSEA (Gene Set Enrichment Analysis)等[50] [51]。通过富集分析的结果，研究人员可以更好地理解基因的功能和调控机制，以及不同基因集合之间的差异。以上为在基因水平上的分析方法，除此之外，还包括基因调控网络推断以及转录因子分析等。

4.2. 单细胞数据常用分析工具

单细胞分析的流程是多个独立开发的工具的集合，这些工具是构建分析流程的基础。随着 scRNA-seq 的发展已经开发了诸多平台，scRNA-seq 分析的工具不断更新和发展。Cell Ranger 是由 10x Genomics 开发的流程化单细胞 RNA 测序分析工具[50]，可以进行数据预处理、转录本比对和定量、细胞识别和聚类等操作。相对于其他工具而言，Cell Ranger 操作简便、速度快，适用于初学者和快速分析。Seurat 是最

受欢迎和最全面的分析平台[52]，可以做 scRNA-seq 数据质量控制、数据分析和数据挖掘，还可以鉴定稀有细胞亚型，最有用的一点就是整合数据分析，它可以利用来自不同测序技术、不同物种和不同处理的数据中共同的变异来进行下游的差异分析。Monocle 则提出了一个在拟时间(Pseudotime)上对单细胞进行排序的策略[53]，利用生物过程中单个细胞之间并不同步的表达状态来还原这个生物过程的细胞轨迹，也可以利于 t-SNE 等降维的方法来进行聚类，然后发现差异表达基因。Scanpy 是一个基于 Python 的不断发展的单细胞分析平台[40]，该平台可以扩展应用于分析更大规模的单细胞数据集，它可以整合以 Python 为编程语言的很多工具，尤其是在机器学习中流行的工具。Scater 在质量控制和数据预处理方面具有特殊优势[54]。其他常用分析平台还包括 TSCAN (Tools for Single-Cell Analysis)以及 RCA (Reference Component Analysis)等[55] [56]，这些分析平台的开发为单细胞数据分析提供了极大的便利。

综上所述，优质的单细胞数据库应该提供完整的数据处理流程和可靠的分析工具，它主要包括：

1) 数据预处理：对原始数据进行质量控制、过滤、去除低质量细胞、去除双峰分布的细胞、批次效应校正等预处理步骤，以保证后续分析的准确性。

2) 细胞类型识别和分类：使用聚类算法对细胞进行分组，识别出不同的细胞类型，并进行分型分析。

3) 基因表达量分析：计算单个基因在每个细胞中的表达量，并进行基因的差异表达分析，以发现不同细胞类型之间的转录差异。

4) 细胞亚型分析：利用细胞的基因表达数据进行亚型分析，发现不同的亚型，并对其进行功能和表型分析。

5) 基因共表达网络分析：使用基因表达数据构建基因共表达网络，识别出共表达模块，并对其进行生物学功能和代谢通路分析，以发现新的生物学模块和代谢通路。

6) 细胞状态识别：通过对不同状态细胞的转录数据进行比较，识别细胞的不同状态，并对其进行功能和表型分析。

7) 数据可视化：通过可视化技术将数据可视化为热图、散点图等形式，以帮助研究者快速了解数据分析结果。

总的来说，优质的单细胞转录组测序数据库应该提供全面的数据分析和可视化功能，以帮助研究者深入理解单个细胞的转录组特征，并为生命科学领域的研究提供有价值的信息和工具。

4.3. 单细胞数据库内置的分析方法及分析工具

目前 scRNA-seq 数据的分析工具数量非常庞大，不同的工具具有不同的特点和功能，单细胞数据库越来越注重对分析方法的使用。例如，HCA 数据库的分析功能较为丰富和全面，可以满足不同用户的需求。从数据可视化方面来看，HCA 数据库提供了丰富的交互式可视化工具，可以帮助用户直观地浏览和分析大规模细胞数据。用户可以通过热图、散点图、堆积图等方式探索不同细胞类型之间的差异和相似性，同时也可以进行空间分析，了解不同细胞在组织中的分布情况。HCA 数据库还提供了一些专门的可视化工具，如“生长轨迹(Trajectory)”和“三维可视化(3D Visualization)”，可帮助用户更好地理解细胞发育和演化的过程。在细胞分类方面，HCA 数据库采用了多种方法来对细胞进行分类，包括传统的基于形态学和生物学特征的分类方法和最新的机器学习方法。这些方法可以帮助用户将不同的细胞类型进行标记和分类，并进行比较和分析。在功能注释方面，HCA 数据库帮助用户分析和注释不同细胞类型的功能和表达谱，包括基因富集分析、网络分析和代谢通路分析等。HCA 数据库使用的分析工具包括 Seurat，Cell Ranger，Scanpy，Monocle 以及 CellProfiler 等。这些工具可以帮助用户深入了解细胞的生物学特征和功能，无论是基础科学研究还是临床应用，都有很大的潜力和应用价值。

HCA 数据库提供了单细胞转录组学数据的资源，而 CancerSEA 数据库属于系统性癌症转录组学分析

的平台。CancerSEA 对带有原始测序文件的 scRNA-seq 数据集采用内部生物信息学管道进行质量控制和表达量化，研究者们根据元数据(Metadata)和 scRNA-seq 推断的拷贝数变异方法去除非恶性单细胞，并过滤了低质量的细胞。CancerSEA 集成了来自 TCGA (The Cancer Genome Atlas)的数千个癌症患者的 RNA-Seq 数据，使用一系列分析工具对这些数据进行标准化、差异表达和聚类分析等。CancerSEA 使用多种计算方法来预测转录因子(Transcription factors, TF)的调控作用，包括基于转录因子结合位点的预测、共表达分析和 TF-TF 互作网络分析。SC2disease 数据库中的条目包含不同细胞类型、组织和疾病相关健康之间差异表达基因的比较[57]，研究者们通过统一管道矩阵重新分析了基因表达，以提高不同研究之间的可比性。CellMarker2.0 数据库中的 Cell Tools 引入了多款单细胞转录组分析软件，包括细胞注释、细胞功能聚类、细胞分化轨迹分析、细胞恶性肿瘤以及细胞通讯等，具体信息如表 2 所示。

Table 2. Methods of database analysis
表 2. 数据库分析的方法

数据库	是否预处理	降维方法	下游分析	分析工具	在线分析功能
SCovid	是	UMAP	Cell annotation, cell clustering, DEG	Seurat, clusterProfiler [58], Cytoscape [59], STRING [60]	无
CancerSEA	是	PCA, t-SNE	Cell clustering	Seurat, DESeq2 [61], GSEA, ConsensusClusterPlus [62], EDGE [63]	无
SC2disease	是	PCA, t-SNE, UMAP, Diffusion Map	Differential expression, gene expression comparison	Seurat, STRING, Cytoscape, Reactome [64], NetworkAnalyst [65]	无
CellMarker2.0	是	UMAP, t-SNE	Cell annotation, cell clustering, cell feature, cell differentiation trajectory analysis, cell malignancy, ell communication	InferCNV, Monocle 3, Seurat	无
Expression Atlas	是	UMAP, t-SNE, PCA	Differential expression analysis, systematic clustering analysis, expression comparison analysis, biological process enrichment analysis, gene co-expression network analysis, gene regulation analysis	Cell Ranger, Seurat, Scanpy, Monocle, CellProfiler [66], GSEA, EnrichR [67], Cytoscape, STRING, BioGRID [68]	无
Cellatlas	是	t-SNE	Cell trajectory, gene regulatory network	Seurat, Scanpy	有
HTCA	是	UMAP, t-SNE, PCA	Data integration, data imputation, dimension reduction, clustering, DE analysis, cell type prediction, manual annotation, data splicing, cell–cell communication	Seurat, Harmony [69], DoubletDecon [70], org.Hs.eg.db, Scasa, SingleR [71]	有
HUSCH	是	UMAP	Functional analyses, transcription regulators, cell–cell interaction analyses, cell type annotation, marker gene identification, differential expression (DE), gene set enrichment analyses (GSEA)	MAESTRO [72]，Harmony, ComplexHeatmap [73], LISA [74], CellChat [75]	无
ABC portal	是	UMAP	Cell annotation, cellular composition, cell-cell communication, gene expression	Cell Ranger, Harmony, DoubletFinder, Seurat, inferCNV, scmap [76], AUCell, CellPhoneDB [77]	无

Continued

Tabula Muris	是	t-SNE, PCA, UMAP	Cell type annotation, differential expression, gene regulatory network analysis, cell state analysis, genealogy analysis, cell subgroup analysis	Seurat, Cell Ranger, SCENIC [78], edgeR, GRNBoost2 [79], Monocle, Scanpy, SingleR	无
HCA	是	t-SNE, PCA, UMAP, Diffusion maps, PHATE	Cell type identification, gene expression analysis, trajectory analysis, subcellular localization analysis, integrated analysis	Seurat, Cell Ranger, Scanpy, Monocle, CellProfiler	无
Allen Cell Types Database	是	t-SNE, PCA, UMAP	Morphological analysis, electrophysiological analysis, gene expression analysis, classification and clustering, molecular analysis, neuron type identification	Neurolucida [80], Clampfit, MATLAB, Allen SDK, CellExplorer	无
The Human BioMolecular Atlas Program (HuBMAP)	是	t-SNE, PCA, ICA, LDA	Cell type identification, subcellular structure analysis, analysis of cell development and differentiation processes, data integration and network analysis	CellProfiler, STAR, Seurat, CellAssign, Cell Ranger, SCENIC, MetaboAnalyst [81], Cytoscape	无
Tabula Muris Senis	是	t-SNE, PCA, UMAP	Cell type annotation, differential expression, functional enrichment analysis	Cell Ranger, Seurat, Scanpy, Mast [82], EdgeR, SingleR, scMatch, CellTypeNet, GO [83], GSEA	无
PanglaoDB	否	/	/	/	无
scRNASeqDB	是	t-SNE, PCA, UMAP	Gene expression differential analysis, enrichment analysis, functional annotation	Seurat, Scanpy, Cell Ranger, Loupe Cell Browser	无
CellBlast	是	t-SNE, PCA, UMAP	Cell type annotation, clue gene analysis, gene expression analysis, cell identification, driver gene analysis	Seurat, Scanpy, Cell Ranger, SC3 [84], CellAssign, SingleR	有

其中，“/”代表原数据库未提供相关功能或者本文未收集到相关内容。

单细胞测序技术的广泛应用已经产生了许多单细胞数据库，这些数据库提供了宝贵的资源来研究生物学中的单细胞异质性和细胞类型。由表 2 可以看出这些数据库通常会进行预处理，包括质量控制、过滤和归一化等，以确保数据的准确性和一致性。主流的单细胞数据降维方法包括 t-SNE、UMAP 和 PCA，这些方法可以将高维数据可视化为二维或三维图像，以便更好地理解和分析数据。在这些数据库中，这些降维方法通常用于绘制聚类图、细胞类型识别和可视化细胞发育轨迹等方面。在单细胞数据库中，主要的数据分析方法包括聚类分析、细胞类型识别、差异表达分析、轨迹分析以及组成成分分析等。这些分析方法可以帮助科学家更好地理解单个细胞和细胞群体之间的差异和联系，而 Seurat、Scanpy 和 Cell Ranger 等分析工具则为我们提供了强大的计算支持。

然而，尽管单细胞数据库提供了丰富的数据资源和多种分析工具，但仍存在一些不足之处。首先，由于不同数据库使用的数据处理和分析方法不同，加上数据文件格式和数据侧重点的不同，数据的对比分析目前仅限于单个数据库内部，这限制了科学家们的研究能力。其次，单细胞数据分析方法和分析工具仅限于常用的几种，无法满足特定研究问题的需求。最后，现有的单细胞数据库大多只对已收集数据进行分析，不提供在线分析功能，这限制了科学家们利用这些数据库进行探索和发现新的生物学知识的

能力。因此，未来的单细胞研究应该致力于建立更加统一和标准化的数据处理和分析方法，并将多个单细胞数据库整合到一起，以便更好地进行数据对比分析。同时，需要不断开发新的分析工具和算法，以满足不同研究问题的需求并开发在线分析功能，以便科学家们更好地利用这些数据库进行研究。

5. 结论与展望

单细胞测序技术彻底改变了我们对细胞异质性的理解，揭示了复杂组织和生物体中以前无法企及的细胞多样性水平。然而，分析和解释单细胞测序产生的大量数据仍然是一个重大挑战，需要开发新的计算和统计学方法。近年来，众多单细胞测序数据库的建立为单细胞测序数据的存储、共享和分析提供了便利。在本文中，我们概述了单细胞测序数据库的现状，讨论了它们的优势、弱点和未来的发展潜力。第一代单细胞测序数据库主要侧重于为原始测序数据提供集中存储库，促进更广泛的科学界重用已发表的数据集。这类数据库的例子包括基因表达综合数据库 GEO (Gene Expression Omnibus)和序列读取存档数据库 SRA (Sequence Read Archive)，它们长期以来一直是批量 RNA-seq 数据的主要存储库。然而，随着单细胞测序领域的成熟，对更专门的数据库的需求，以适应单细胞数据的独特要求变得越来越明显。第二代单细胞测序数据库包括单细胞门户 SCP (Single Cell Portal)和单细胞表达图谱 SCEA (Single Cell Expression Atlas)等资源，它们提供了更复杂的数据处理和分析管道，使用户能够更详细地探索和可视化单细胞测序数据。这些数据库通常包括用于质量控制、标准化和细胞聚类的工具，以及用于方便识别细胞类型和状态的交互式可视化和探索工具。

单细胞测序领域面临的关键挑战之一是不同技术和平台产生的数据的整合。鉴于技术的快速发展，新的测序平台和方案正在不断开发。为了应对这一挑战，已经启动了几项计划，旨在为单细胞测序数据开发标准化数据格式和元数据模式。例如，HCA 项目已经为单细胞测序数据的生成、处理和共享制定了一套数据标准和协议，目标是创建所有人类细胞的全面图谱。

在开发单细胞测序数据库时，另一个重要因素是平衡数据可访问性与数据隐私和安全性。单细胞测序数据具有固有的敏感性，包含关于组织或生物内单个细胞的身份和特征的信息。因此，单细胞测序数据库必须纳入适当的数据安全和隐私措施，如去识别和访问控制，以确保数据的使用符合伦理和负责任。

此外，目前的单细胞转录组测序数据库仍然缺乏对于除人类和小鼠等以外物种的关注，标记基因等重要信息也比较稀缺。单个数据库数据内容关注点过于片面，没有全面涵盖单细胞转录组测序研究中的多个方面。其次，单细胞数据库中用于单细胞数据分析的工具屈指可数，单细胞文件格式不一，在线分析功能的建设也有待完善。

为了解决上述问题，未来单细胞数据库建设可以关注以下几点：

1) 增加数据库的数据量并扩大关注面。在现有单细胞转录组测序数据库的基础上，可以增加对于其他物种的数据收集和处理。另外，可以加强标记基因等重要信息的收集和整理，以满足单细胞转录组测序研究的需求。

2) 加强单细胞数据的分析算法的开发。在单细胞转录组测序数据库中，用于单细胞数据分析的工具数量仍然较少，分析功能也相对简单。未来的单细胞转录组测序研究需要关注分析算法的开发，提高数据分析的准确性和效率。

3) 需要关注对于不同格式数据文件的标准化处理，设定全面统一的分析管道，提高不同数据库的连通性及数据的可比性。这将有助于更好地整合和比较不同数据库中的数据，推进单细胞转录组测序研究的跨数据库整合和比较分析。

单细胞转录组测序研究是一个快速发展的领域，随着数据量和技术水平的不断提高，我们相信这一领域将会在细胞生物学、疾病诊断和治疗等方面取得更大的突破和进展。未来单细胞转录组测序数据库

的建设需要关注多个方面的需求，以更好地推动单细胞转录组测序研究的发展。

基金项目

本研究由国家自然科学基金(项目编号：12090052)提供资助。

参考文献

[1] Hwang, B., Lee, J.H. and Bang, D. (2018) Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines. *Experimental & Molecular Medicine*, **50**, 1-14. https://doi.org/10.1038/s12276-018-0071-8

[2] Buettner, F., *et al.* (2015) Computational Analysis of Cell-to-Cell Heterogeneity in Single-Cell RNA-Sequencing Data Reveals Hidden Subpopulations of Cells. *Nature Biotechnology*, **33**, 155-160. https://doi.org/10.1038/nbt.3102

[3] Hu, H., *et al.* (2022) CITEMO^{XMBD}: A Flexible Single-Cell Multimodal Omics Analysis Framework to Reveal the Heterogeneity of Immune Cells. *RNA Biology*, **19**, 290-304. https://doi.org/10.1080/15476286.2022.2027151

[4] Shimizu, H. and Nakayama, K.I. (2020) Artificial Intelligence in Oncology. *Cancer Science*, **111**, 1452-1460. https://doi.org/10.1111/cas.14377

[5] Kaufmann, S.H.E. (2019) Immunology's Coming of Age. *Frontiers in Immunology*, **10**, Article 684. https://doi.org/10.3389/fimmu.2019.00684

[6] Wu, W., *et al.* (2022) Exploring the Cellular Landscape of Circular RNAs Using Full-Length Single-Cell RNA Sequencing. *Nature Communications*, **13**, Article No. 3242. https://doi.org/10.1038/s41467-022-30963-8

[7] Kim, D., *et al.* (2020) The Architecture of SARS-CoV-2 Transcriptome. *Cell*, **181**, 914-921. https://doi.org/10.1016/j.cell.2020.04.011

[8] Mathys, H., *et al.* (2019) Single-Cell Transcriptomic Analysis of Alzheimer's Disease. *Nature*, **570**, 332-337. https://doi.org/10.1038/s41586-019-1195-2

[9] Giustacchini, A., *et al.* (2017) Single-Cell Transcriptomics Uncovers Distinct Molecular Signatures of Stem Cells in Chronic Myeloid Leukemia. *Nature Medicine*, **23**, 692-702. https://doi.org/10.1038/nm.4336

[10] Segerstolpe, A., *et al.* (2016) Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, **24**, 593-607. https://doi.org/10.1016/j.cmet.2016.08.020

[11] Whiteside, T.L. (2008) The Tumor Microenvironment and Its Role in Promoting Tumor Growth. *Oncogene*, **27**, 5904-5912. https://doi.org/10.1038/onc.2008.271

[12] Janakiraman, M., *et al.* (2010) Genomic and Biological Characterization of Exon 4 KRAS Mutations in Human Cancer. *Cancer Research*, **70**, 5901-5911. https://doi.org/10.1158/0008-5472.CAN-10-0192

[13] Barretina, J., *et al.* (2012) The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature*, **483**, 603-607. https://doi.org/10.1038/nature11003

[14] Muus, C., *et al.* (2021) Single-Cell Meta-Analysis of SARS-CoV-2 Entry Genes across Tissues and Demographics. *Nature Medicine*, **27**, 546-559. https://doi.org/10.1038/s41591-020-01227-z

[15] Zhang, J.-Y., *et al.* (2020) Single-Cell Landscape of Immunological Responses in Patients with COVID-19. *Nature Immunology*, **21**, 1107-1118. https://doi.org/10.1038/s41590-020-0762-x

[16] Liao, M., *et al.* (2020) Single-Cell Landscape of Bronchoalveolar Immune Cells in Patients with COVID-19. *Nature Medicine*, **26**, 842-844. https://doi.org/10.1038/s41591-020-0901-9

[17] Wilk, A.J., *et al.* (2020) A Single-Cell Atlas of the Peripheral Immune Response in Patients with Severe COVID-19. *Nature Medicine*, **26**, 1070-1076. https://doi.org/10.1038/s41591-020-0944-y

[18] Hu, H., *et al.* (2023) Modeling and Analyzing Single-Cell Multimodal Data with Deep Parametric Inference. *Briefings in Bioinformatics*, **24**, Article No. bbad005. https://doi.org/10.1093/bib/bbad005

[19] Qi, C., *et al.* (2022) SCovid: Single-Cell Atlases for Exposing Molecular Characteristics of COVID-19 across 10 Human Tissues. *Nucleic Acids Research*, **50**, D867-D874. https://doi.org/10.1093/nar/gkab881

[20] Yuan, H., *et al.* (2019) CancerSEA: A Cancer Single-Cell State Atlas. *Nucleic Acids Research*, **47**, D900-D908. https://doi.org/10.1093/nar/gky939

[21] Natarajan, K.N., *et al.* (2019) Comparative Analysis of Sequencing Technologies for Single-Cell Transcriptomics. *Genome Biology*, **20**, Article No. 70. https://doi.org/10.1186/s13059-019-1676-5

[22] Hu, H., *et al.* (2023) Gene Function and Cell Surface Protein Association Analysis Based on Single-Cell Multiomics Data. *Computers in Biology and Medicine*, **157**, Article ID: 106733. https://doi.org/10.1016/j.compbiomed.2023.106733

[23] Zhang, X., *et al*. (2019) CellMarker: A Manually Curated Resource of Cell Markers in Human and Mouse. *Nucleic Acids Research*, **47**, D721-D728. https://doi.org/10.1093/nar/gky900

[24] Hu, C., *et al*. (2023) CellMarker 2.0: An Updated Database of Manually Curated Cell Markers in Human/Mouse and Web Tools Based on scRNA-Seq Data. *Nucleic Acids Research*, **51**, D870-D876. https://doi.org/10.1093/nar/gkac947

[25] Papatheodorou, I., *et al*. (2020) Expression Atlas Update: From Tissues to Single Cells. *Nucleic Acids Research*, **48**, D77-D83.

[26] Zeng, J., *et al*. (2022) CancerSCEM: A Database of Single-Cell Expression Map across Various Human Cancers. *Nucleic Acids Research*, **50**, D1147-D1155. https://doi.org/10.1093/nar/gkab905

[27] Wang, R., *et al*. (2023) Construction of a Cross-Species Cell Landscape at Single-Cell Level. *Nucleic Acids Research*, **51**, 501-516. https://doi.org/10.1093/nar/gkac633

[28] Pan, L., *et al*. (2023) HTCA: A Database with an In-Depth Characterization of the Single-Cell Human Transcriptome. *Nucleic Acids Research*, **51**, D1019-D1028. https://doi.org/10.1093/nar/gkac791

[29] Shi, X., *et al*. (2023) HUSCH: An Integrated Single-Cell Transcriptome Atlas for Human Tissue Gene Expression Visualization and Analyses. *Nucleic Acids Research*, **51**, D1029-D1037. https://doi.org/10.1093/nar/gkac1001

[30] Gao, X., *et al*. (2023) ABC Portal: A Single-Cell Database and Web Server for Blood Cells. *Nucleic Acids Research*, **51**, D792-D804. https://doi.org/10.1093/nar/gkac646

[31] Han, X., *et al*. (2018) Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, **172**, 1091-1107. https://doi.org/10.1016/j.cell.2018.02.001

[32] Schaum, N., *et al*. (2018) Single-Cell Transcriptomics of 20 Mouse Organs Creates a *Tabula Muris*. *Nature*, **562**, 367-372. https://doi.org/10.1038/s41586-018-0590-4

[33] Tasic, B., *et al*. (2016) Adult Mouse Cortical Cell Taxonomy Revealed by Single Cell Transcriptomics. *Nature Neuroscience*, **19**, 335-346. https://doi.org/10.1038/nn.4216

[34] Lake, B.B., *et al*. (2018) Integrative Single-Cell Analysis of Transcriptional and Epigenetic States in the Human Adult Brain. *Nature Biotechnology*, **36**, 70-80. https://doi.org/10.1038/nbt.4038

[35] The Tabula Muris Consortium (2020) A Single-Cell Transcriptomic Atlas Characterizes Ageing Tissues in the Mouse. *Nature*, **583**, 590-595. https://doi.org/10.1038/s41586-020-2496-1

[36] Cao, Y., Zhu, J., Han, G., Jia, P. and Zhao, Z. (2017) ScRNASeqDB: A Database for Gene Expression Profiling in Human Single Cell by RNA-Seq. BioRxiv, Article ID: 104810. https://doi.org/10.1101/104810

[37] Franzén, O., Gan, L.-M. and Björkegren, J.L.M. (2019) PanglaoDB: A Web Server for Exploration of Mouse and Human Single-Cell RNA Sequencing Data. *Database*, **2019**, Article No. baz046. https://doi.org/10.1093/database/baz046

[38] Pardi, N., Hogan, M.J., Porter, F.W. and Weissman, D. (2018) MRNA Vaccines—A New Era in Vaccinology. *Nature Reviews Drug Discovery*, **17**, 261-279. https://doi.org/10.1038/nrd.2017.243

[39] Liu, Y., Beyer, A. and Aebersold, R. (2016) On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*, **165**, 535-550. https://doi.org/10.1016/j.cell.2016.03.014

[40] Wolf, F.A., Angerer, P. and Theis, F.J. (2018) SCANPY: Large-Scale Single-Cell Gene Expression Data Analysis. *Genome Biology*, **19**, Article No. 15. https://doi.org/10.1186/s13059-017-1382-0

[41] Becht, E., *et al*. (2019) Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nature Biotechnology*, **37**, 38-44. https://doi.org/10.1038/nbt.4314

[42] Kodinariya, T.M. and Makwana, P.R. (2013) Review on Determining Number of Cluster in K-Means Clustering. *International Journal*, **1**, 90-95.

[43] Murtagh, F. and Contreras, P. (2017) Algorithms for Hierarchical Clustering: An Overview, II. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, **7**, e1219. https://doi.org/10.1002/widm.1219

[44] Bhattacharjee, P. and Mitra, P. (2021) A Survey of Density Based Clustering Algorithms. *Frontiers of Computer Science*, **15**, Article No. 151308. https://doi.org/10.1007/s11704-019-9059-3

[45] Schaeffer, S.E. (2007) Graph Clustering. *Computer Science Review*, **1**, 27-64. https://doi.org/10.1016/j.cosrev.2007.05.001

[46] Saelens, W., Cannoodt, R., Todorov, H. and Saeys, Y. (2019) A Comparison of Single-Cell Trajectory Inference Methods. *Nature Biotechnology*, **37**, 547-554. https://doi.org/10.1038/s41587-019-0071-9

[47] Peng, L., *et al*. (2022) Cell-Cell Communication Inference and Analysis in the Tumour Microenvironments from Single-Cell Transcriptomics: Data Resources and Computational Strategies. *Briefings in Bioinformatics*, **23**, Article No. bbac234. https://doi.org/10.1093/bib/bbac234

[48] Ji, Q., *et al*. (2019) Single-Cell RNA-Seq Analysis Reveals the Progression of Human Osteoarthritis. *Annals of the*

Rheumatic Diseases, **78**, 100-110. https://doi.org/10.1136/annrheumdis-2017-212863

[49] Scialdone, A., *et al*. (2015) Computational Assignment of Cell-Cycle Stage from Single-Cell Transcriptome Data. *Methods*, **85**, 54-61. https://doi.org/10.1016/j.ymeth.2015.06.021

[50] Zheng, G.X.Y., *et al*. (2017) Massively Parallel Digital Transcriptional Profiling of Single Cells. *Nature Communications*, **8**, Article ID: 14049.

[51] Subramanian, A., *et al*. (2005) Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550. https://pubmed.ncbi.nlm.nih.gov/16199517/

[52] Butler, A., Hoffman, P., Smibert, P., Papalexi, E. and Satija, R. (2018) Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species. *Nature Biotechnology*, **36**, 411-420. https://doi.org/10.1038/nbt.4096

[53] Trapnell, C., *et al*. (2014) The Dynamics and Regulators of Cell Fate Decisions Are Revealed by Pseudotemporal Ordering of Single Cells. *Nature Biotechnology*, **32**, 381-386. https://doi.org/10.1038/nbt.2859

[54] McCarthy, D.J., Campbell, K.R., Lun, A.T. and Wills, Q.F. (2017) Scater: Pre-Processing, Quality Control, Normalization and Visualization of Single-Cell RNA-Seq Data in R. *Bioinformatics*, **33**, 1179-1186. https://doi.org/10.1093/bioinformatics/btw777

[55] Ji, Z. and Ji, H. (2016) TSCAN: Pseudo-Time Reconstruction and Evaluation in Single-Cell RNA-Seq Analysis. *Nucleic Acids Research*, **44**, e117. https://doi.org/10.1093/nar/gkw430

[56] Li, H., *et al*. (2017) Reference Component Analysis of Single-Cell Transcriptomes Elucidates Cellular Heterogeneity in Human Colorectal Tumors. *Nature Genetics*, **49**, 708-718. https://doi.org/10.1038/ng.3818

[57] Zhao, T., *et al*. (2021) SC2disease: A Manually Curated Database of Single-Cell Transcriptome for Human Diseases. *Nucleic Acids Research*, **49**, D1413-D1419. https://doi.org/10.1093/nar/gkaa838

[58] Yu, G., Wang, L.-G., Han, Y. and He, Q.-Y. (2012) ClusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters. *OMICS*: *A Journal of Integrative Biology*, **16**, 284-287. https://doi.org/10.1089/omi.2011.0118

[59] Shannon, P., *et al*. (2003) Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, **13**, 2498-2504. https://doi.org/10.1101/gr.1239303

[60] Szklarczyk, D., *et al*. (2019) STRING v11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Research*, **47**, D607-D613. https://doi.org/10.1093/nar/gky1131

[61] Love, M.I., Huber, W. and Anders, S. (2014) Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biology*, **15**, Article No. 550. https://doi.org/10.1186/s13059-014-0550-8

[62] Wilkerson, M.D. and Hayes, D.N. (2010) ConsensusClusterPlus: A Class Discovery Tool with Confidence Assessments and Item Tracking. *Bioinformatics*, **26**, 1572-1573. https://doi.org/10.1093/bioinformatics/btq170

[63] Storey, J.D. and Tibshirani, R. (2003) Statistical Significance for Genomewide Studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440-9445. https://doi.org/10.1073/pnas.1530509100

[64] Jassal, B., *et al*. (2020) The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, **48**, D498-D503.

[65] Xia, J., Benner, M.J. and Hancock, R.E. (2014) NetworkAnalyst—Integrative Approaches for Protein-Protein Interaction Network Analysis and Visual Exploration. *Nucleic Acids Research*, **42**, W167-W174. https://doi.org/10.1093/nar/gku443

[66] Carpenter, A.E., *et al*. (2006) CellProfiler: Image Analysis Software for Identifying and Quantifying Cell Phenotypes. *Genome Biology*, **7**, Article No. R100. https://doi.org/10.1186/gb-2006-7-10-r100

[67] Kuleshov, M.V., *et al*. (2016) Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update. *Nucleic Acids Research*, **44**, W90-W97. https://doi.org/10.1093/nar/gkw377

[68] Chatr-Aryamontri, A., *et al*. (2017) The BioGRID Interaction Database: 2017 Update. *Nucleic Acids Research*, **45**, D369-D379. https://doi.org/10.1093/nar/gkw1102

[69] Korsunsky, I., *et al*. (2019) Fast, Sensitive and Accurate Integration of Single-Cell Data with Harmony. *Nature Methods*, **16**, 1289-1296. https://doi.org/10.1038/s41592-019-0619-0

[70] DePasquale, E.A.K., *et al*. (2019) DoubletDecon: Deconvoluting Doublets from Single-Cell RNA-Sequencing Data. *Cell Reports*, **29**, 1718-1727. https://doi.org/10.1016/j.celrep.2019.09.082

[71] Aran, D., *et al*. (2019) Reference-Based Analysis of Lung Single-Cell Sequencing Reveals a Transitional Profibrotic Macrophage. *Nature Immunology*, **20**, 163-172. https://doi.org/10.1038/s41590-018-0276-y

[72] Li, Y., Shi, W. and Wasserman, W.W. (2018) Genome-Wide Prediction of Cis-Regulatory Regions Using Supervised Deep Learning Methods. *BMC Bioinformatics*, **19**, Article No. 202. https://doi.org/10.1186/s12859-018-2187-1

[73] Gu, Z., Eils, R. and Schlesner, M. (2016) Complex Heatmaps Reveal Patterns and Correlations in Multidimensional Genomic Data. *Bioinformatics*, **32**, 2847-2849. https://doi.org/10.1093/bioinformatics/btw313

[74] Rousseeuw, P.J. (1987) Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

[75] Jin, S., *et al*. (2021) Inference and Analysis of Cell-Cell Communication Using CellChat. *Nature Communications*, **12**, Article No. 1088. https://doi.org/10.1038/s41467-021-21246-9

[76] Kiselev, V.Y., Yiu, A. and Hemberg, M. (2018) Scmap: Projection of Single-Cell RNA-Seq Data across Data Sets. *Nature Methods*, **15**, 359-362. https://doi.org/10.1038/nmeth.4644

[77] Efremova, M., Vento-Tormo, M., Teichmann, S.A. and Vento-Tormo, R. (2020) CellPhoneDB: Inferring Cell-Cell Communication from Combined Expression of Multi-Subunit Ligand-Receptor Complexes. *Nature Protocols*, **15**, 1484-1506. https://doi.org/10.1038/s41596-020-0292-x

[78] Baron, M., *et al*. (2016) A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Systems*, **3**, 346-360. https://doi.org/10.1016/j.cels.2016.08.011

[79] Moerman, T., *et al*. (2019) GRNBoost2 and Arboreto: Efficient and Scalable Inference of Gene Regulatory Networks. *Bioinformatics*, **35**, 2159-2161. https://doi.org/10.1093/bioinformatics/bty916

[80] Ghahremani, P., *et al*. (2021) NeuroConstruct: 3D Reconstruction and Visualization of Neurites in Optical Microscopy Brain Images. *IEEE Transactions on Visualization and Computer Graphics*, **28**, 4951-4965. https://doi.org/10.1109/TVCG.2021.3109460

[81] Chong, J., Yamamoto, M. and Xia, J. (2019) MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. *Metabolites*, **9**, Article No. 57. https://doi.org/10.3390/metabo9030057

[82] Finak, G., *et al*. (2015) MAST: A Flexible Statistical Framework for Assessing Transcriptional Changes and Characterizing Heterogeneity in Single-Cell RNA Sequencing Data. *Genome Biology*, **16**, Article No. 278. https://doi.org/10.1186/s13059-015-0844-5

[83] Ashburner, M., *et al*. (2000) Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, **25**, 25-29. https://doi.org/10.1038/75556

[84] Kiselev, V.Y., *et al*. (2017) SC3: Consensus Clustering of Single-Cell RNA-Seq Data. *Nature Methods*, **14**, 483-486. https://doi.org/10.1038/nmeth.4236