Biophysics 生物物理学, 2022, 10(2), 31-38
Published Online May 2022 in Hans. http://www.hanspub.org/journal/biphy
https://doi.org/10.12677/biphy.2022.102004

Hans汉斯

基于自然语言处理的单细胞转录组数据伪时间分析

卢雨儿[1,2]，胡　桓[1,2]，陈玲玲[1]，程　烽[1]，帅建伟[1,2]，林　海[1,2*]

[1]厦门大学物理系，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：2022年4月22日；录用日期：2022年5月22日；发布日期：2022年5月31日

摘　要

针对单细胞转录组测序数据，人们已经提出了各种强大的分析模型和处理算法，用于细胞聚类、细胞类型识别、细胞伪时间轨迹推断、细胞RNA动力学、基因调控网络推断和RNA速度分析等。本文提出一种方法，将自然语言处理技术引入单细胞转录组数据分析中。算法首先采用TF-IDF表示转录组基因表达强度对细胞功能的影响程度，进一步把细胞演化发育过程所形成的各种基因表达变化，理解为自然语言中的各种句子文本，创新性地把自然语言文本分析技术应用于单细胞转录组演化发育的处理。通过在基因网络上随机行走生成各种基因序列文本，从而生成基因空间中基因的嵌入式词向量表示和细胞的嵌入式词向量表示，实现了对单细胞转录组数据的伪时间可视化分析。最后的分析结果表明该模型对于单细胞数据进行细胞发育伪时间分析是一种有效的方法。

关键词

单细胞测序，伪时间轨迹推断，自然语言处理，基因组学

Pseudo-Time Analysis of Single-Cell Transcriptome Data Based on Natural Language Processing

Yu'er Lu[1,2], Huan Hu[1,2], Lingling Chen[1], Feng Cheng[1], Jianwei Shuai[1,2], Hai Lin[1,2*]

[1]Department of Physics, Xiamen University, Xiamen Fujian
[2]Wenzhou Institute, University of Chinese Academy of Sciences, Wenzhou Zhejiang

Received: Apr. 22nd, 2022; accepted: May 22nd, 2022; published: May 31st, 2022

*通讯作者。

文章引用: 卢雨儿, 胡桓, 陈玲玲, 程烽, 帅建伟, 林海. 基于自然语言处理的单细胞转录组数据伪时间分析[J]. 生物物理学, 2022, 10(2): 31-38. DOI: 10.12677/biphy.2022.102004

Abstract

For single-cell transcriptome sequencing data, various powerful analytical models and processing algorithms have been proposed for cell clustering, cell type recognition, cell pseudo-time trajectory inference, cellular RNA dynamics, gene regulatory network inference, and RNA velocity analysis. This paper proposes an innovative approach to introducing natural language processing techniques into single-cell transcriptome data analysis. The algorithm first uses TF-IDF to indicate the degree of influence of transcriptome gene expression intensity on cell function, and further innovatively treats the various gene expression changes formed by the process of cell evolution and development as various sentence texts in natural language. Then, the natural language text analysis can be applied for the processing of evolutionary development of single-cell transcriptomes. Various gene sequence texts are generated by random walking process on the gene network, which generates the embedded word vector representation of genes and the embedded word vector representation of cells in the gene space, respectively. Finally, the pseudo-time visual analysis is considered for the single-cell transcriptome data. The final analysis results show that this model is an effective method for pseudo-time analysis of cell development for single-cell data.

Keywords

Single-Cell Sequencing, Pseudo-Time Trajectory Inference, Natural Language Processing, Genomics

1. 引言

2009 年，汤富酬在剑桥大学 Gurdon 研究所攻读博士后，作为第一作者在《Nature Methods》发表了世界上第一篇单细胞 mRNA 测序文章[1]，提出了一种单细胞全转录组扩增方法，进行第一例单细胞 RNA 测序，分析了来自小鼠四细胞胚胎阶段的单个卵裂球的转录组，开创了单细胞测序技术[1]。单细胞测序技术可谓是科技发展史上的一大创举，它极大地推进了基因组学领域的研究，使不同细胞类型得以精细区分，使科学家能在单细胞水平上进行分子机制的研究[2]。单细胞测序技术旨在单细胞水平揭示细胞的基因结构和基因转录组表达状态，反映细胞间基因表达的异质性，为细胞分子生命科学的研究提供独特的视角。

十多年来，单细胞测序方法蓬勃发展，这些发展使单细胞测序具有通量高、周期快、成本低、细胞捕获率高等优点。现在单细胞测序技术流程包括：首先将单个细胞进行分离，并确保其生物完整性不被破坏。目前常用的单细胞分离方法有连续稀释法、激光捕获显微切割术、显微操作法、荧光激活细胞分选术、拉曼镊子技术和微流控技术等。然后是细胞溶解与基因组获取，对细胞进行溶解来获取 DNA 或 RNA 基因组，该关键步骤技术的挑战是尽量保证基因组的完整性。目前细胞溶解的方法可以分为物理法、化学法和生物酶降解法。接着进行全基因组扩增，由于单个细胞中基因含量无法达到测序仪的检测线，因此需要对基因组进行扩增，目前常用方法是利用 DNA 聚合酶和不同形式的引物来进行扩增，引物包括特异性的、简并的或杂合的引物。近年来，单细胞测序分析在肿瘤、发育生物学、微生物学、神经科学等领域发挥着日益重要的作用，成为了生命科学极具潜力的热点研究工具[3] [4] [5] [6]。

一般的单细胞转录组数据分析流程，首先是对单细胞 RNA 测序平台给出的原始数据进行预处理，包括数据质控、数据标准化、数据缺失填充、数据去除过滤、批次效应处理等，从而得到单细胞基因表达矩阵。然后采用各种降维和可视化方法，对单细胞基因表达矩阵进行分析处理，包括细胞亚群分类、各亚群差异表达基因分析、标记基因筛选等分析。在此基础上，可进一步进行伪时间轨迹、细胞通讯、转录调控网络、RNA 速度等各种研究分析[7] [8]。由于单细胞测序技术能够快速确定成千上万个细胞的精确基因表达模式，从而分析相同表型细胞的遗传信息异质性，目前已大量应用于器官生长、胚胎学、神经生物学、免疫学、微生物学、产前基因诊断、癌症生物学、临床医疗诊断等多个研究领域[9]-[15]。

在过去的十多年中，已经开发了各种不同的单细胞伪时间分析方法[16] [17] [18] [19] [20]。在本文，我们提出了一种基于自然语言处理的伪时间分析算法，首先提出了一种结合 TF-IDF 和余弦相似度的细胞间距离表示方法，进一步将单细胞转录组数据进行词嵌入文本处理，再通过可视化展示细胞轨迹并推断出单细胞伪时间排列，从而能够对细胞分化轨迹进行重建，为研究细胞动态分化过程提供了一个新的分析方法。

2. 模型算法

我们的伪时间轨迹分析模型主要包含 4 个模块，如图 1 所示，分别是单细胞数据预处理模块、基因相似性网络构建模块、细胞文本语言处理模块、伪时间分析模块。下面我们将分别对该模型四个模块进行详细介绍。

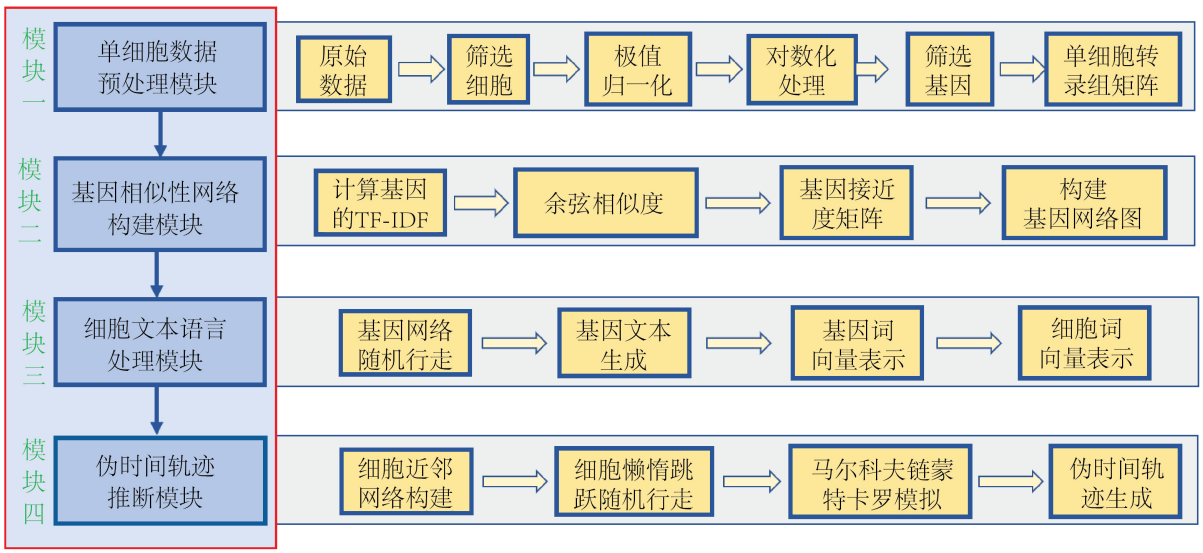

Figure 1. Workflow of pseudo-time trajectory model based on natural language processing
图 1. 基于自然语言处理的伪时间轨迹模型流程图

2.1. 数据预处理

基于数据特征所展示的单细胞测序技术局限所带来的影响，我们将对单细胞数据进行一些预处理。首先对细胞原始数据进行质量控制，去掉低质量的细胞，然后对转录组表达进行归一化处理和对数化处理，得到处理后的单细胞转录组矩阵。

2.2. 基因相似性网络构建

在细胞这个生命基本单元中，基因支持着生命的基本构造和性能，通常一个细胞中包含上万种基因，

这些基因对细胞功能也有不同的贡献。受到自然语言处理中 TF-IDF 技术的启发[21] [22]，我们认为，这些不同贡献可以由基因的重要性来表示，因此，为表示一个细胞中不同基因对细胞功能的不同贡献，我们定义了基因的词频－逆词频指数 $TFIDF_{ij}$，该指数能够表示基因 j 对细胞 i 功能的重要程度，即当某个基因 j 在某个细胞 i 中的 $TFIDF$ 较大时，就认为这个基因对这个细胞较为重要。

为此，首先定义基因频率：

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

其中，n_{ij} 是细胞 i 中基因 j 的表达量，分母是基因 j 在所有细胞中的表达量。

其次定义基因 j 的逆文件频率：

$$IDF_j = \log\frac{|C|}{1+\left|\left(k : g_j \in c_k\right)\right|}$$

其中，$|C|$ 表示细胞总数，$\left|\left(k : g_j \in c_k\right)\right|$ 表示包含包含基因 j 的细胞数量。然后得到细胞 i 中基因 j 的 $TFIDF$ 值：

$$TFIDF_{ij} = TF_{ij} \times IDF_j$$

最后，基于余弦相似度定义基因 i 和基因 j 之间的接近度：

$$\varnothing_{ij} = \cos\theta = \frac{\sum_k G_{ik} \times G_{jk}}{\sqrt{G_i^2} \times \sqrt{G_j^2}}$$

其中，G_{ik} 表示第 k 个细胞中第 i 个基因的 $TFIDF$ 值。

我们将基因看成节点，基因和基因之间的接近度作为节点之间连接的权重，构成网络图。基因节点构成的集合 $V = \{v_1, \cdots, v_n\}$，节点之间的连接的边由 E 表示，边的长度由权重 W 表示，W 是一个 $n \times n$ 的矩阵，描述了基因 i 与基因 j 之间的接近度 w_{ij}。这个矩阵称为邻接矩阵。取一个合适的阈值，基因之间的相似度大于该阈值时，认为这两个基因之间有边相连，就构成基因的网络图。

2.3. 细胞文本语言处理

基于基因网络图，我们采用随机行走的方式来构建基因文本。在网络中取初始节点，可遍历取点，也可随机取点，从初始节点开始随机行走，即从一个节点以边权重为概率走到下一个邻居节点。

对一个无向连通图，$w_{ij} = w_{ji}$，概率转移矩阵 P，在图 G 上的标准随机游走由下式给出

$$P = D^{-1}W$$

其中 D 是 $n \times n$ 的矩阵，是每个节点的度加权和的对角矩阵，矩阵元素表示为

$$d_{ij} = \begin{cases} \sum_k w_{ik}, & i = j \\ 0, & i \neq j \end{cases}$$

随机行走后记录走过的每一个基因，即形成由基因序列构成的基因文本。

接下来，我们采用词嵌入算法[23]，将基因序列构成的基因文本转换成基因的词向量。词嵌入过程就是把一个维数为所有词数量的高维空间嵌入到一个维数低的连续向量空间中，每个单词或词组被映射为实数域上的向量，词嵌入的结果就生成了词向量。获得基因的向量表示后，由于每个单细胞表达一组基因，将单细胞表达的基因矢量以表达量为权重加和，形成的合矢量可作为该细胞在基因空间中的矢量表

示：

$$C_{mk} = \sum_i a_{mi} g_{ik}$$

其中 C_{mk} 代表第 m 个细胞矢量的第 k 维， a_{mi} 代表第 m 个细胞矢量的第 i 个基因的表达量， g_{ik} 代表第 i 个基因矢量的第 k 维。

2.4. 细胞伪时间分析

最后一个是单细胞伪时间轨迹推断模块。这个模块我们采用 VIA 算法[24]。构建好细胞的向量表示后，可以计算细胞之间的欧几里得距离，得到细胞的距离矩阵。在细胞距离矩阵的基础上，先通过聚类将细胞降维，然后在降维后的细胞网络中进行“懒惰－跳跃随机行走”，从而得到细胞网络的起始节点和终止分枝节点，再通过马尔科夫链蒙特卡罗模拟，计算最可能的行走路线，最后得到细胞的伪时间轨迹。

2.5. 单细胞转录组数据

本研究涉及的数据集是人类胚胎干单细胞转录组测序数据集，简称为 EB [25]。该数据集包含了在 27 天(5 个时间段)的分化过程中的 31000 个人类胚胎细胞数据，覆盖了 33694 种基因表达。这些细胞均匀分布在整个胚胎分化过程中，它们分别被命名为 0~3 d，6~9 d，12~15 d，18~21 d 和 24~27 d。

3. 分析结果

图 2 是人类胚胎干细胞转录组测序数据集小提琴图，左图展示的是细胞内基因的种类数,中间图展示的是细胞的基因表达总量,右图表示的线粒体基因表达量在所有基因表达总量的占比。对此我们将进行预处理操作，先对数据进行质量控制，筛选出高质量的检测细胞数据，尽可能地去除低质量细胞，例如死细胞，油滴未捕获的空包等。其次是归一化处理，减小基因因为表达量级不同带来的下游分析影响和消除特殊基因表达细胞数据导致的不良影响，接下来进行取根号操作，以平滑数据分布。

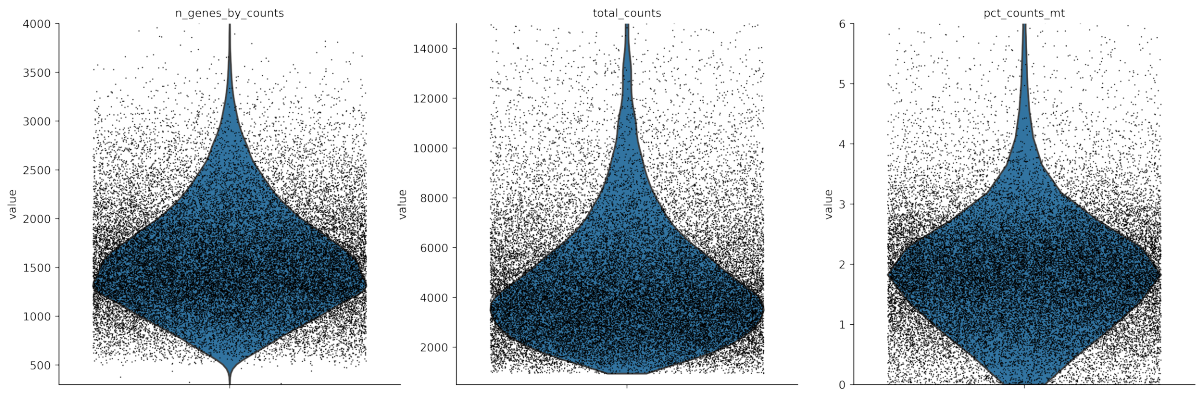

Figure 2. Violin plots of human embryonic stem cell transcriptome sequencing data, including the number of gene species in cells (left), the total amount of gene expression in cells (middle), and the percentage of mitochondrial gene expression in the total amount of gene expression (right)

图 2. 人类胚胎干细胞转录组数据小提琴图，包括细胞内基因种类数(左)，细胞的基因表达总量(中)，和线粒体基因表达量在所有基因表达总量的占比(右)

预处理后先计算每个细胞中每个基因的 TF-IDF 值，然后基于余弦距离计算基因和基因之间的相似度，得到基因接近度矩阵。将接近度看成邻接矩阵，就得到了基因网络。为了便于可视化网络结构，我们取

一个合适的阈值 0.5，去掉权重低于该阈值的连接，计算该网络的最小支撑树，可以得到如图 3(a)所展示的人类胚胎干细胞基因网络结构图，网络中的一个节点代表一种基因，节点间连线代表为基因之间的关联。在基因网络图上进行随机行走，生成基因文本，再通过词嵌入算法将其编码为基因的词向量表示。图 3(b)是用 UMAP 算法得到的关于基因的矢量表示图，图中的每个点代表人类胚胎细胞中的一种基因。

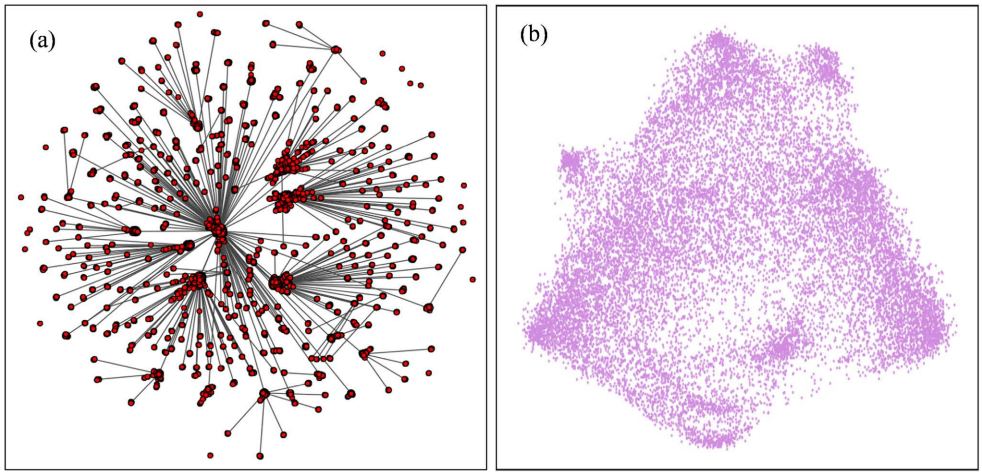

Figure 3. Human embryonic stem cell gene network structure (a) and gene vector representation (b)
图 3. 人类胚胎干细胞基因网络结构图(a)和基因矢量表示图(b)

获得胚胎细胞基因的词向量表示后，由于每个单细胞表达一组基因，因此我们将胚胎细胞表达的基因矢量以表达量为权重加和形成合矢量，该矢量可作为该细胞在基因空间中的词向量表示。图 4 是通过 UMAP 降维算法，将细胞的矢量表示呈现在二维空间进行可视化结果图。从图 4 可以看出，基于 UMAP 算法的可视化结果图，展示出了人体胚胎干细胞从第一天到第五天的一个不断发育的过程，这个可以从图中颜色从红色到蓝色的过渡看出。除此之外，图中从红色到蓝色存在细胞分化产生分支的过程，这说明我们的模型可以很好地展示细胞分化发育的过程。

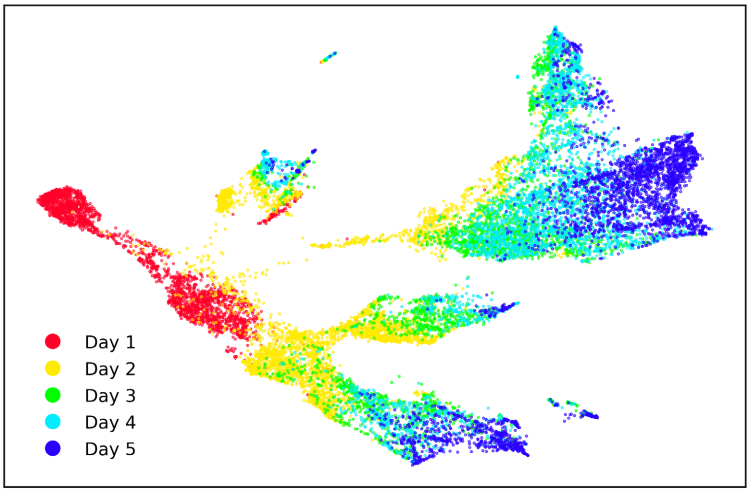

Figure 4. Pseudo-time distribution of human embryonic stem cells based on UMAP
图 4. 基于 UMAP 的人类胚胎干单细胞伪时间分布图

最后，我们采用前述的 VIA 算法[24]，计算细胞发育分化的轨迹，结果如图 5 所示。可以看到，该轨迹也能较好地反映了细胞发育的线路、方向及起始和终末节点。从第一天红色标示的起始细胞开始，第二天黄色标示的细胞主要出现两个分叉线路，到第三和第四天出现更多分叉线路，最后第五天的蓝色细胞呈现出各种的终末节点群。

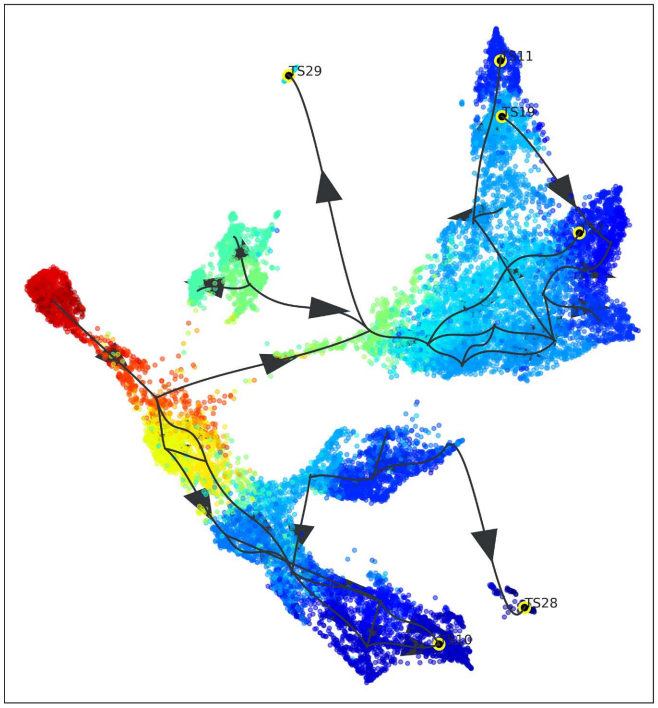

Figure 5. Pseudo-time trajectory inference of human embryonic stem cells based on VIA algorithm
图 5. 基于 VIA 的人类胚胎干单细胞伪时间轨迹推断

4. 结论

本文提出了一种基于自然语言处理的单细胞分化伪时间分析算法。不同于现在的各种关于单细胞伪时间轨迹的人工智能和深度学方法分析方法，我们的模型创新性地结合自然语言处理技术中的 TF-IDF 及余弦相似度的概念，得到了基因接近度矩阵，从而构建了基因网络图。进一步，我们把细胞演化发育过程所形成的各种基因变异轨迹，理解为自然语言中的各种句子文本，从而首次将自然语言处理技术引入伪时间轨迹分析中。我们定义了基因嵌入式词向量表示和细胞嵌入式词向量表示，从而构建了细胞近邻网络。在此基础上，采用 UMAP 降维算法，实现了单细胞伪时间分布的可视化构建，再通过 VIA 算法，得到了完整的轨迹线路图。

本文提出的基于自然语言处理分析单细胞伪时间轨迹模型，在重建细胞分化轨迹的分析上有一定成果。然而，由于单细胞测序数据的复杂性和细胞分化的异质性，该模型仍有很大的修改完善空间。例如，在本模型基础上，进一步发展更好的轨迹推断算法、处理细胞数上万的单细胞数据集等挑战问题，都值得我们进一步去发展讨论。

基金项目

本论文获得国家自然科学基金项目(批准号：11874309，12090052)的资助。

参考文献

[1] Tang, F., Barbacioruet, C., Wang, Y., *et al*. (2009) mRNA-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat Methods*, **6**, 377-382. https://doi.org/10.1038/nmeth.1315

[2] Owens, B. (2012) Genomics: The Single Life. *Nature*, **491**, 27-29. https://doi.org/10.1038/491027a

[3] Potter, S.S. (2018) Single-Cell RNA Sequencing for the Study of Development, Physiology and Disease. *Nature Reviews Nephrology*, **14**, 479-492. https://doi.org/10.1038/s41581-018-0021-7

[4] Baslan, T. and Hicks, J. (2017) Unravelling Biology and Shifting Paradigms in Cancer with Single-Cell Sequencing. *Nature Reviews Cancer*, **17**, 557-569. https://doi.org/10.1038/nrc.2017.58

[5] Kester, L. and van Oudenaarden, A. (2018) Single-Cell Transcriptomics Meets Lineage Tracing. *Cell Stem Cell*, **23**, 166-179. https://doi.org/10.1016/j.stem.2018.04.014

[6] Papalexi, E. and Satija, R. (2018) Single-Cell RNA Sequencing to Explore Immune Cell Heterogeneity. *Nature Reviews Immunology*, **18**, 35-45. https://doi.org/10.1038/nri.2017.76

[7] Carter, B. and Zhao, K. (2021) The Epigenetic Basis of Cellular Heterogeneity. *Nature Reviews Genetics*, **22**, 235-250. https://doi.org/10.1038/s41576-020-00300-0

[8] Woyke, T., D.F.R. Doud, and F. Schulz (2017) The Trajectory of Microbial Single-Cell Sequencing. *Nature Methods*, **14**, 1045-1054. https://doi.org/10.1038/nmeth.4469

[9] Sade-Feldman, M., Yizhak, K., Nordman, E., *et al*. (2018) Defining T Cell States Associated with Response to Checkpoint Immunotherapy in Melanoma. *Cell*, **175**, 998-1013.e20. https://doi.org/10.1016/j.cell.2018.10.038

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

[11] Su, Y., Chen, D., Yuan, D., *et al*. (2020) Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. *Cell*, **183**, 1479-1495.e20. https://doi.org/10.1016/j.cell.2020.10.037

[12] Maier, B., Leader, A.M., Chen, S.T., *et al*. (2020) A Conserved Dendritic-Cell Regulatory Program Limits Antitumour Immunity. *Nature*, **580**, 257-262. https://doi.org/10.1038/s41586-020-2134-y

[13] Bocchi, V.D., Conforti, P., Vezzoli, E., *et al*. (2021) The Coding and Long Noncoding Single-Cell Atlas of the Developing Human Fetal Striatum. *Science*, **372**, Article No. abf5759. https://doi.org/10.1126/science.abf5759

[14] Bhaduri, A., Sandoval-Espinosa, C., Otero-Garcia, M., *et al*. (2021) An Atlas of Cortical Arealization Identifies Dynamic Molecular Signatures. *Nature*, **598**, 200-204. https://doi.org/10.1038/s41586-021-03910-8

[15] Hu, H., Liu, R., Zhao, C., *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

[16] 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

[17] Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F. and Theis, F.J. (2016) Diffusion Pseudotime Robustly Reconstructs Lineage Branching. *Nature Methods*, **13**, 845-848. https://doi.org/10.1038/nmeth.3971

[18] Setty, M., Tadmor, M.D., Reich-Zeliger, S., *et al*. (2016) Wishbone Identifies Bifurcating Developmental Trajectories from Single-Cell Data. *Nature Biotechnology*, **34**, 637-645. https://doi.org/10.1038/nbt.3569

[19] Qiu, X., Mao, Q., Tang, Y., *et al*. (2017) Reversed Graph Embedding Resolves Complex Single-Cell Trajectories. *Nature Methods*, **14**, 979-982. https://doi.org/10.1038/nmeth.4402

[20] Setty, M., Kiseliovas, V., Levine, J., Gayoso, A., Mazutis, L. and Pe'er, D. (2019) Characterization of Cell Fate Probabilities in Single-Cell Data with Palantir. *Nature Biotechnology*, **37**, 451-460. https://doi.org/10.1038/s41587-019-0068-4

[21] Cong, Y., Chan, Y.B. and Ragan, M.A. (2016) Exploring Lateral Genetic Transfer among Microbial Genomes Using TF-IDF. *Scientific Reports*, **6**, Article No. 29319. https://doi.org/10.1038/srep29319

[22] Moussa, M. and Mandoiu, I.I. (2018) Single Cell RNA-seq Data Clustering Using TF-IDF Based Methods. *BMC Genomics*, **19**, Article No. 569. https://doi.org/10.1186/s12864-018-4922-4

[23] Wu, F., Zhang, C. and Zhang, L. (2021) A Deep Learning Framework Combined with Word Embedding to Identify DNA Replication Origins. *Scientific Reports*, **11**, Article No. 844. https://doi.org/10.1038/s41598-020-80670-x

[24] Stassen, S.V., Yip, G.G.K., Wong, K.K.Y., Ho, J.W.K. and Tsia, K.K. (2021) Generalized and Scalable Trajectory Inference in Single-Cell Omics Data with VIA. *Nature Communications*, **12**, Article No. 5528. https://doi.org/10.1038/s41467-021-25773-3

[25] Moon, K.R., van Dijk, D., Wang, Z., *et al*. (2019) Visualizing Structure and Transitions in High-Dimensional Biological Data. *Nature Biotechnology*, **37**, 1482-1492. https://doi.org/10.1038/s41587-019-0336-3