
科学家实现多项单细胞多组学分析算法的系统性评估

作者：writer 来源：中国科学院

本文原地址：<https://www.iikx.com/news/progress/29763.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

科学家实现多项单细胞多组学分析算法的系统性评估。

近日，中国科学技术大学生命科学与医学部教授瞿昆课题组、数学科学学院教授陈发来课题组以及北京生命科学研究所研究员黎斌课题组合作完成了一项成果。该团队分析了百万量级单细胞多组学数据，系统评估了14种单细胞模态预测算法和18种单细胞多组学整合算法的性能。相关研究成果以Benchmarking algorithms for single-cell multi-omics prediction and integration为题，在线发表在《自然-方法》（Nature Methods）上。

单细胞多组学技术的发展，为探索细胞功能和复杂的基因调控机制提供了机遇。然而，湿实验方法通常面临着高成本、数据质量有限及批次效应等挑战。为克服这些局限，生物信息学家基于统计模型和人工智能技术，开发了多种算法。这些算法能够利用单细胞转录组数据推断同一细胞内的蛋白质丰度和染色质可及性信息，并通过将不同模态的数据映射到统一的特征空间实现数据整合以去除批次效应。这些工具提升了现有单细胞数据的解析能力。然而，面对海量数据和众多算法，科研人员往往难以判断哪些工具最适合他们的研究工作。因此，对这些算法进行基准测试尤为重要。

该研究收集了47个数据集的上百万个单细胞多组学数据，涵盖多个生物样本和实验平台。研究设计了一套全面的评估流程，结合算法的准确性、鲁棒性和计算资源消耗等多维度指标，评估了领域内的常用算法。结果显示，在蛋白质丰度预测方面，totalVI和scArches表现最优异；在染色质可及性预测中，LS_Lab算法排名领先；在多组学整合分析中，Seurat、MOJITOO和scAI在垂直整合上表现突出，而totalVI和UINMF在水平整合和马赛克整合任务中展现出优异性能。该工作为算法设计提供了新思路，并为未来多组学数据的分析和应用奠定了基础。同时，为帮助科研人员选择合适的分析工具，该团队在GitHub上发布了完整的分析流程、代码和测试数据集，以供同行使用和改进。

进一步，科研人员探讨了这些算法的数学原理，发现了降噪处理是提高单细胞数据预测精度的关键。在性能评估中，机器学习算法以及基于概率模型的深度学习算法表现出优势。同时，研究提出，现有模态预测算法在某些关键蛋白的预测性能上有待提升，染色质可及性预测的准确性需要进一步优化。

研究工作得到国家重点研发计划和国家自然科学基金等的资助，并获得中国科大超级计算中心与生命科学学院生物信息学中心的计算资源支持。

[论文链接](#)

评估流程示意图

研究团队单位：中国科学技术大学

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归原作者所有，请勿用于商业用途，[爱科学iikx.com](#)转发