# QB期刊 陈盛泉团队对单细胞DNA甲基化数据的填补策略进行了综合评估

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/37352.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！*

**QB期刊 陈盛泉团队对单细胞DNA甲基化数据的填补策略进行了综合评估**。论文标题：Imputing not available values in single-cell DNA methylation data using the median is straightforward and effective

期刊：Quantitative Biology

作者：Songming Tang, Siyu Li, Shengquan Chen

发表时间：5 Feb 2025

DOI： 10.1002/qub2.70000

微信链接：点击此处阅读微信文章

DNA甲基化（DNAm）是最早被发现的表观遗传修饰类型之一，且在调控正常细胞周期、胚胎发生以及肿瘤发展进程中扮演着重要角色。近年来，单细胞DNA甲基化（scDNAm）的进展为以极高分辨率探索细胞表观遗传差异提供了前所未有的机会。当前大多数研究分析单细胞DNA甲基化数据时，通常基于细胞—区域矩阵。创建scDNAm数据细胞—区域矩阵的简单有效方法是基因组窗口分箱，即将基因组划分为固定长度的区块（例如100 kbp），计算每个细胞在每个区域中的平均DNA甲基化水平。然而，在进行后续分析之前，仍需解决一个关键问题：如何处理scDNAm数据中的缺失值（NA）。对于单细胞RNA测序（scRNA-seq）或单细胞转座酶可及染色质测序（scATAC-seq）数据，测序中的缺失通常表现为零读数。然而，在scDNAm数据中，捕获的甲基化位点通常具有二元特征：甲基化（读数为1）或非甲基化（读数为0），而未捕获的位点则标记为NA。在使用窗口分箱策略构建细胞—区域矩阵时，由于甲基化位点在基因组中的分布不均以及窗口大小的影响，许多区域可能没有捕获到甲基化位点，导致其平均甲基化水平标记为NA值（即缺失值）。具有NA值的甲基化矩阵无法进行后续分析，因此，填补缺失值成为必要的预处理步骤。

近期，南开大学数学科学学院陈盛泉课题组在Quantitative Biology期刊发表了一篇题目名为Imputing not available values in single-cell DNA methylation data using the median is straightforward and effective的文章，揭示了利用中位数填充单细胞DNA甲基化数据是一种简单有效的方式。

Received: 24 October 2024 | Accepted: 5 February 2025
DOI: 10.1002/qub2.70000

COMMUNICATION

Quantitative Biology
Open Access

# Imputing not available values in single-cell DNA methylation data using the median is straightforward and effective

Songming Tang | Siyu Li | Shengquan Chen

全文概要

在分析scDNAm数据时，一种直观的解决方案是将所有缺失值（NA）填补为零。然而，从另一个角度来看，单细胞RNA测序（scRNA-seq）中较高的读数通常对应于较高的基因表达水平，而基因表达与DNA甲基化水平呈显著负相关。因此，scRNA-seq数据中的缺失值通常被视为零，相当于将scDNAm数据中的NA值填补为一。此外，利用各种统计方法对NA值进行平滑处理是一种直观的做法。例如，EpiScanpy使用区域内所有细胞的甲基化水平均值来填补NA值。该研究发现，使用中位数填补NA值能更有效地填补scDNAm数据中的缺失值，并改善后续分析流程。

聚类和可视化性能评估

研究对11个不同来源、协议和物种的数据集进行了综合测试，以评估不同填补策略对scDNAm数据的影响（图1）。研究首先针对数据中的CpG甲基化生成了区域长度为100 kbp的细胞—区域矩阵，应用了四种不同的填补策略：0、1、均值和中位数对缺失值进行填补，并使用主成分分析方法将数据降至50维，利用Louvain算法进行无监督细胞聚类，并使用统一流形近似和投影方法可视化从不同填补方法获得的嵌入数据。聚类和可视化的结果表明，中位数填补策略始终提供了稳健且优异的表现。

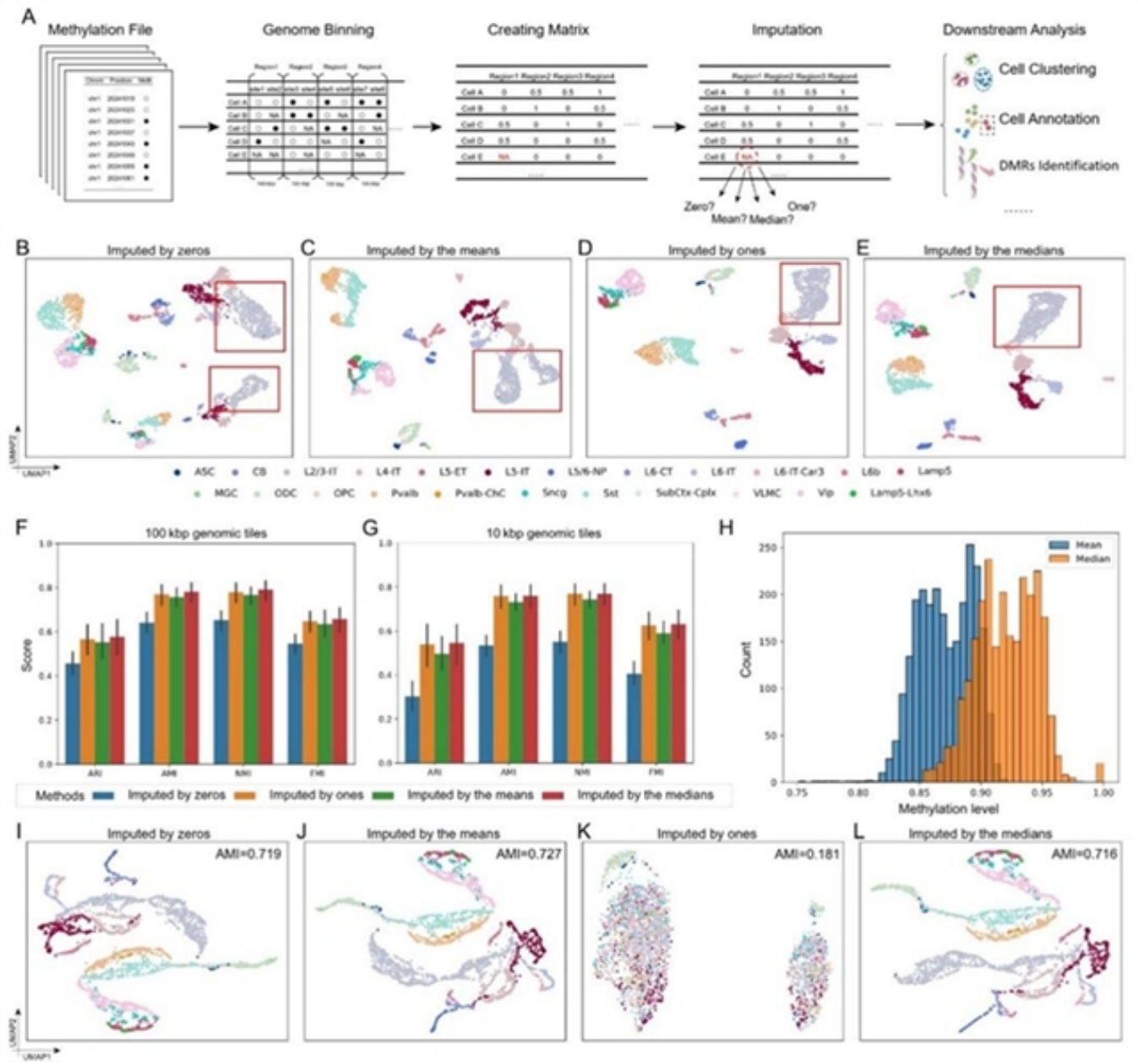



图1.不同填补策略的效果

在更高分辨率下的评估

为了在更高区域分辨率下研究填补策略，研究人员将区域长度降至10 kbp，并重复上述实验。结果表明，在较短区域下，中位数和1填补几乎等效，并且对NA值的处理非常有效。与100 kbp区域相比，10 kbp区域的缺失数据问题更加严重，大多数可用值显著偏离零。因此，0填补在较短的区域长度中极不适用，因为在缺失数据增加的情况下，0填补会引入更多噪声。

针对CH甲基化的评估

对于CH甲基化数据，研究人员使用100 kbp区域长度进行了实验。中位数填补的CH甲基化数据结果始终表现良好。然而，由于CH位点的甲基化比例极低，使用1填补引入了显著噪声，导致在

填补数据中丧失了细胞异质性。

总结

总结来说，该项研究建议使用中位数填补NA值，这是一种简单有效的方法，有助于突出单细胞DNA甲基化数据中的细胞异质性，为后续分析提供准确的数据基础，并允许对潜在生物学过程做出更精确可靠的解释。

QB期刊介绍

Quantitative Biology （QB）期刊是由清华大学、北京大学、高教出版社联合创办的全英文学术期刊。QB主要刊登生物信息学、计算生物学、系统生物学、理论生物学和合成生物学的最新研究成果和前沿进展，并为生命科学与计算机、数学、物理等交叉研究领域打造一个学术水平高、可读性强、具有全球影响力的交叉学科期刊品牌。

《前沿》系列英文学术期刊

由教育部主管、高等教育出版社主办的《前沿》（Frontiers）系列英文学术期刊，于2006年正式创刊，以网络版和印刷版向全球发行。系列期刊包括基础科学、生命科学、工程技术和人文社会科学四个主题，是我国覆盖学科最广泛的英文学术期刊群，其中12种被SCI收录，其他也被AHCI、Ei、MEDLINE或相应学科国际权威检索系统收录，具有一定的国际学术影响力。系列期刊采用在线优先出版方式，保证文章以最快速度发表。

中国学术前沿期刊网

http://journal.hep.com.cn




来源：Quantitative Biology

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