scATAC 数据模拟

自硕 21 崔晏菲 2021210976

1. 背景

ATAC-seq（Assay for Transposase-Accessible Chromatin with highthroughput sequencing）是利用转座酶研究染色质开放性的高通量测序技术。在此基础上，scATAC-seq 通过 barcode 识别细胞核，以完成对细胞核的分选和标记，解决了不同细胞群体的异质性的问题，从而能够检测出混杂样品测序所无法得到的异质性信息。因此 scATAC-seq 是一种表征细胞类型特异性调控环境的强大方法。在 2015 年，Buenrostro 等人[1]就利用 scATAC 测序技术揭示了转录景观。后来，人们也发明了许多 scATAC-seq 的数据分析方法[2-7]。

然而，这些方法都存在一个共同的问题：由于我们没有对于细胞类型的先验知识的金标准，所以很难评估这些数据分析技术的功效[8]。那么从另外一条路出发，为了评估这些方法，我们也可以用真实标签来模拟 scATAC-seq 数据，通过模拟，我们就可以将不同的细胞类型转化为各式各样的可量化的参数，这样就可以对这些数据分析技术进行评估分析。

2. 研究进展

2.1 基于采样的模拟方法

目前已经有很多人都在研究模拟 scATAC-seq 数据，起初最简单的方法就是从精心挑选的 bulk ATAC-seq 数据中进行下采样[3]，

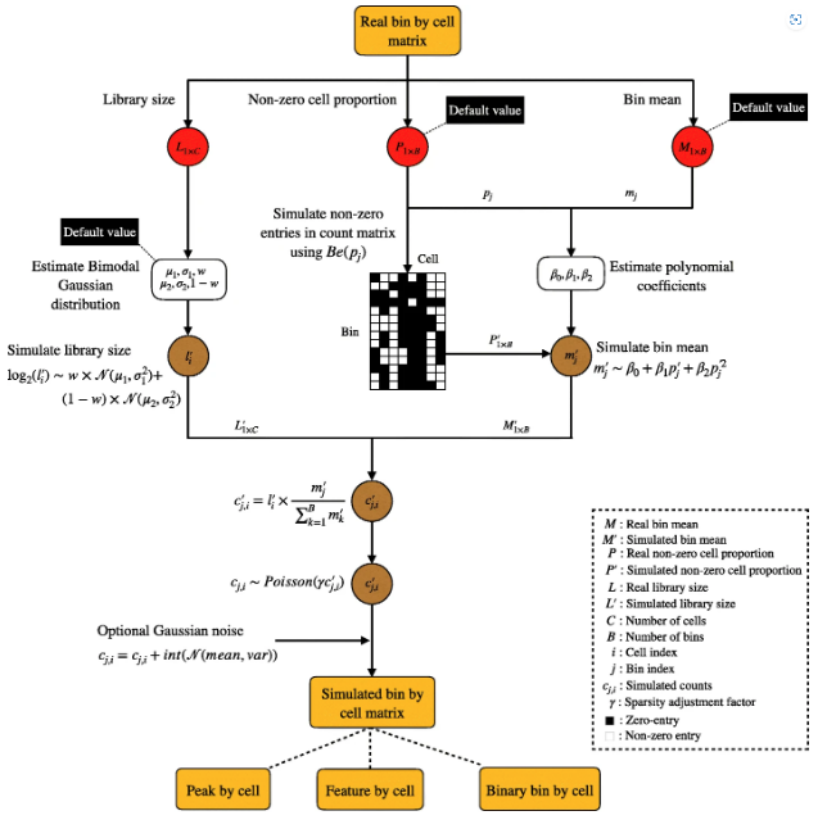
这个方法较为简单，但是信噪比相当低，引入了批次效应的偏差，还违反了 scATAC-seq 实验的二倍体性质。为了解决这些问题，人们又选择具有严格固定信噪比的前景峰采样来模拟数据[6, 9]，然而这样的采样又忽视了 scATAC-seq 数据的背景区域。基于如此，人们发明了 SCAN-ATAC-Sim 方法进行模拟[8]，首先对于每个细胞，人们对它的前景和背景区域进行两次采样，并为每个采样区域随机选择一个读数，接着将来自前景和背景的读数组合在一起以形成最终的读数。相比于前两种方法来说，这个方法的进步是显著的，不仅增大了信噪比，而且同时考虑到了前景区域和背景区域，还大大提高了运算速度。以上三种方法的对比如下图所示：

Features	Direct down sampling	Peak region sampling	SCAN-ATAC Sim
Read-level simulation	Yes	No	Yes
Flexible signal-to-noise	No	No	Yes
Diploid genomic constraint	No	Yes	Yes
Short runtime	No	No	Yes
Long flag	Default value	Long flag	Default value
–cell_number	10 000	–min_frag	1000
–signal_to_noise	0.7	–max_frag	20 000
–frag_num	3000	–extend_peak_size	1000
–variance	0.5	–bin_size	1000

2.2 基于特征矩阵的模拟方法

然而，基于采样的模拟方法仍然存在许多问题，一方面 scATAC-seq 数据具有稀疏性和噪声性，这对采样方法来说是极易产生误差的；另一方面，在 scATAC-seq 的数据处理中，很重要的一步是生成 scATAC-seq 数据的特征矩阵，因此，基于特征矩阵进行 scATAC-seq 数据的模拟是一个可行的思路。

人们提出了 simATAC 方法[10]，给定一个真实的 scATAC-seq 特征矩阵作为输入，simATAC 就可以按细胞类型来估计读数分布的统计参数，并生成一个合成读数阵列，该阵列可以捕获具有相似生物学特征的细胞的独特调控景观。实验证明 simATAC 模拟的数据和真实数据具有极大的相似性。下图是 simATAC 的流程示意图：



simATAC 的优点有三个：第一，这是一个在离散、连续和伪真实方面综合模拟 scATAC-seq 的模型；第二，它的统计数据和真实数据在峰值、均值、稀疏度方面非常一致；第三，它通过用户自定义的单元格实现了灵活的仿真。

然而，simATAC 仍然有很多不足有待改进，首先就是它的仿真依然受到批次效应的影响很大，其次对峰值、均值还有稀疏度等指标只是做了简单的对比，并没有在统计学意义上做建模证明。但这仍

然是一个非常不错的方法，这是迈向生成式模型的一大步。

3. 参考文献

1. Buenrostro, J.D., et al., *Single-cell chromatin accessibility reveals principles of regulatory variation.* Nature, 2015. **523**(7561): p. 486-490.
2. Bravo González-Blas, C., et al., *cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data.* Nature methods, 2019. **16**(5): p. 397-400.
3. Fang, R., et al., *SnapATAC: A comprehensive analysis package for single cell ATAC-seq.* BioRxiv, 2020: p. 615179.
4. Liu, L., et al., *Deconvolution of single-cell multi-omics layers reveals regulatory heterogeneity. Nat Commun 10 (1): 470*. 2019.
5. Schep, A.N., et al., *chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data.* Nature methods, 2017. **14**(10): p. 975-978.
6. Xiong, L., et al., *Scale method for single-cell atac-seq analysis via latent feature extraction. Nat. Commun, 10 (4576)*. 2019.
7. Zamanighomi, M., et al., *Unsupervised clustering and epigenetic classification of single cells. Nat. Commun, 9 (2410)*. 2018.
8. Chen, Z., et al., *SCAN-ATAC-Sim: a scalable and efficient method for simulating single-cell ATAC-seq data from bulk-tissue experiments.* Bioinformatics, 2021. **37**(12): p. 1756-1758.

9. Zhang, J., et al., *An integrative ENCODE resource for cancer genomics.* Nature communications, 2020. **11**(1): p. 1-11.

10. Navidi, Z., L. Zhang, and B. Wang, *simATAC: a single-cell ATAC-seq simulation framework.* Genome Biology, 2021. **22**(1).