中国科学: 数学 2025 年 第 55 卷 第 7 期: 1383 ~ 1398
SCIENTIA SINICA Mathematica
综 述

单细胞扰动数据分析

献给钱敏平教授 86 寿辰

付子初[1], 杨茗[1], 侯琳[1,2*]

1. 清华大学统计与数据科学系, 北京 100084;
2. 清华大学生命科学学院生物信息学教育部重点实验室, 北京 100084
E-mail: fzc21@mails.tsinghua.edu.cn, ming-yang@mail.tsinghua.edu.cn, houl@tsinghua.edu.cn

收稿日期: 2024-10-13; 接受日期: 2025-02-07; 网络出版日期: 2025-05-06; * 通信作者
国家重点研发计划 (批准号: 2020YFA0712403) 和国家自然科学基金 (批准号: T2322017) 资助项目

摘要 单细胞扰动技术将基因编辑或药物处理等外部扰动与单细胞测序结合起来, 提供了在分子层面上细胞对扰动响应的高分辨率表型特征. 通过计算方法对这些数据进行分析有助于揭示基因调控网络和药物作用机制, 并预测未观测基因或药物组合的潜在效应. 本文总结具有代表性的单细胞扰动技术, 概述基于统计建模、机器学习和深度学习等多种扰动效应的解析和预测方法, 并展望扰动图谱、空间多组学数据和因果学习等前沿技术与方法的应用前景.

关键词 扰动 单细胞测序 机器学习 深度学习

MSC (2020) 主题分类 62P10, 92B20, 68T09

1 引言

单细胞扰动筛选技术 (single-cell perturbation screening) 作为现代细胞生物学研究的重要工具, 在理解细胞状态的多样性及其调控机制方面发挥着关键作用 (参见文献 [20,45,65,82]). 细胞处于动态变化的表型空间中, 其状态受内部基因调控与外部环境因素的共同影响 (参见文献 [50,68,72]). 通过对细胞施加特定扰动 (如利用 CRISPR-Cas9 技术进行基因编辑), 研究人员能够揭示细胞潜在的多种状态及其相互转化的路径 (参见文献 [82]). 虽然并非所有状态都可以互相转换, 例如成熟的上皮细胞无法直接转变为免疫细胞, 但通过外部扰动可以揭示其潜在的表型变化范围 (参见文献 [23]).

为了更深入地探索复杂的表型空间, 扰动筛选技术至关重要. 通过在基因层面 (如基因敲除 [15]、过表达 [45] 和干扰 [71]) 或非基因层面 (如抗体 [5]、化合物 [65] 和细胞因子 [34]) 对细胞进行精准扰动, 研究人员能够测量扰动后细胞的表型特征, 并探索细胞状态之间的转换机制. 然而, 传统筛选技术由于只能测量单一或少数表型特征, 限制了我们对表型空间的全面理解. 此外, 由于潜在的扰动间相互组合数量庞大, 传统方法难以全面描绘这一复杂的调控网络和表型空间.

英文引用格式: Fu Z C, Yang M, Hou L. A review on single-cell perturbation data analysis (in Chinese). Sci Sin Math, 2025, 55: 1383–1398, doi: 10.1360/SSM-2024-0315

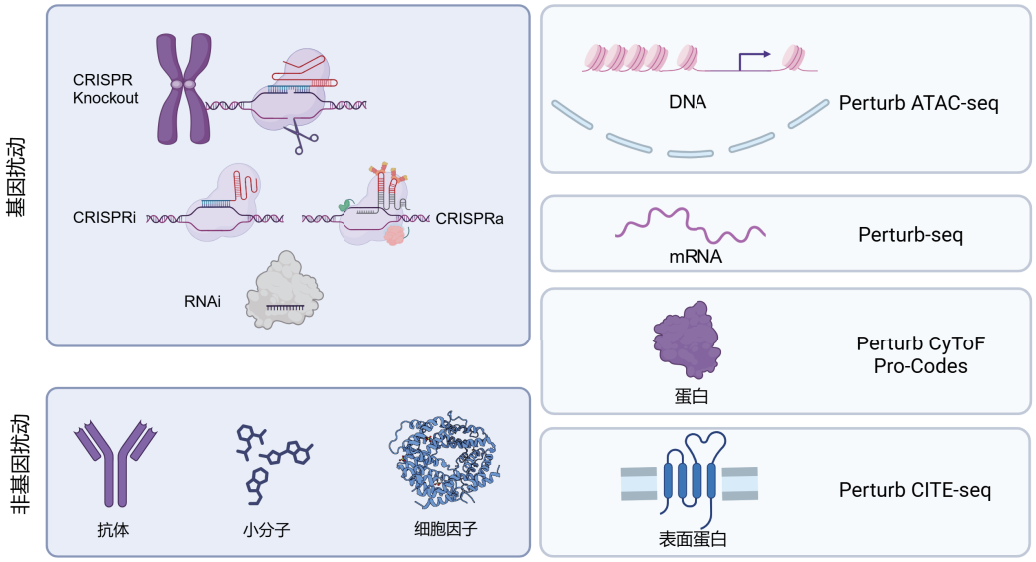

图 1 (网络版彩图) 单细胞扰动类别及检出技术示意图. 目前单细胞扰动技术包含的扰动包括直接扰动基因组的 CRISPR-Cas9、激活靶基因转录的 CRISPRa、阻断靶基因转录的 CRISPRi、切割靶 mRNA 并促进其降解的 RNAi、抗体 (antibodies)、结合细胞表面受体的细胞因子 (cytokines) 以及各类化学小分子等. 这些不同的扰动作用于基因表达和蛋白质生产的不同阶段. 单细胞测序技术用于检测细胞对这些扰动的反应, 覆盖中心法则的不同阶段: scATAC-seq 用于探测染色质状态, scRNA-seq 用于测量 mRNA 表达水平, 而 Pro-Code 用于检测扰动后的蛋白水平, CITE-seq 则通过测量与蛋白质结合的抗体来确定表面蛋白质计数

近年来, 实验技术和计算方法的不断进步正在突破这些局限. 在实验技术层面, 单细胞扰动筛选技术的发展使研究人员能够在单细胞水平上获得复杂的高维表型信息. 这些技术将扰动与单细胞测序 (single-cell sequencing) 结合, 能够捕捉单个细胞的扰动信息, 并记录细胞的染色质开放状态 [58]、基因表达 [15] 和表面蛋白水平 [19] 等丰富的分子特征 (见图 1). 此外, 随着实验规模的扩大, 研究人员可以同时处理成千上万个扰动及其组合, 为全面描绘细胞状态提供了前所未有的可能.

在计算方法层面, 各种计算工具的发展推动了对细胞内调控机制的理解, 有助于构建能够分析与预测细胞状态及功能的模型 (参见文献 [21,32,67]). 单细胞扰动数据的计算模型主要围绕两个核心目标: 解析和预测 [55]. 针对基因层面的扰动, 解析方法有助于揭示扰动对基因 [78]、基因模块 [17]、细胞全局状态的影响 [4], 分析扰动相似性 [60], 探索基因之间的相互作用 [45]; 而预测方法则可以推测实验未观测到的基因或其组合的扰动效应 [56], 并通过主动学习或强化学习优化实验设计, 进一步探索尚未充分了解的表型空间 [28]. 对于非基因层面的扰动, 解析方法在药物再利用 [62,70]、揭示药物作用机制 [59] 等方面具有重要意义; 预测方法则支持虚拟药物筛选与多种药物联合疗法效果的预测 [32,38], 从而为新药开发和疾病治疗策略提供有力支持.

综上所述, 单细胞扰动技术使我们能够更系统地描绘细胞表型图谱, 加深对细胞生物学机制的理解, 同时为疾病的诊断与治疗开辟新的路径. 本综述讨论该领域中的代表性技术、现有数据类型及其分析方法, 并指出未来研究中的关键挑战与发展前景.

2 单细胞扰动技术与数据

本节对具有代表性的单细胞扰动技术进行简要介绍, 这些方法的实验步骤和具体细节的比较可以

参阅近期的综述 [8].

Perturb-seq [15] 和 CRISP-seq [31] 是单细胞扰动领域的初期突破性技术, 它们的核心在于利用 CRISPR 系统 (CRISPR pooled screens) 进行基因编辑, 通过在单细胞中引入特定的单一引导 RNA (single guide RNA, sgRNA) 以干扰目标基因的功能. 每个 sgRNA 编码质粒中都附有独特的引导条形码 (guide barcode), 从而在测序过程中可以追踪每个细胞内的 sgRNA 类型. 在实验中, 这两项技术首先将带有 sgRNA 的质粒导入细胞, 然后通过逆转录过程将细胞内的内源性 mRNA 和带有条形码的 sgRNA 转录本在同一细胞条形码 (cell barcode) 下进行连接, 随后进行高通量测序. 这种方法能够同时捕获每个细胞的基因表达数据和施加的基因干扰信息, 从而全面揭示基因功能和细胞反应. 然而, 这些技术也存在一些限制和挑战, 例如, 在逆转录过程中, 可能会出现模板切换效应 (template-switching effect), 尤其是在病毒共包装时, 这可能导致错误的条形码关联, 从而影响数据的准确性. 此外, 受实验条件或技术因素的影响, sgRNA 的干扰效果可能不完全或不一致, 从而影响基因功能的准确分析.

CROP-seq [12] 是另一项广泛应用的技术, 其核心创新在于能够直接读取单细胞中 sgRNA 的信息. 在 CROP-seq 中, 每个 sgRNA 被设计成带有聚腺苷酸化尾巴的形式, 允许在单细胞中进行转录本的直接捕获和测序. 这使得该技术能够处理大规模的引导 RNA 文库, 并且与 CRISPR 筛选的标准克隆协议兼容, 从而提升了 CRISPR 筛选的效率和精确度. CROP-seq 也面临一些挑战, 如 sgRNA 的独特设计使得载体长度有所限制, 这可能会影响组合扰动的效果和 sgRNA 的检索效率. Mosaic-seq [77] 侧重于解析表观遗传学调控及其对基因表达的影响. 该技术利用 CRISPR/dCas9 系统引导表观遗传调控因子靶向特定基因位点, 从而在单细胞水平上引入特定的表观遗传修饰, 以研究这些修饰对基因表达和细胞功能的影响. Mosaic-seq 的主要优势在于其能够在单细胞层面高分辨率地解析表观遗传调控网络, 为研究基因表达的调控机制提供了有力的工具.

除了靶向基因位点外, 一些技术可以得到药物扰动后细胞的分子特征图谱. MIX-seq [41] 是一项基于单核苷酸多态性 (single-nucleotide polymorphism, SNP) 进行细胞群体分辨 (demultiplexing) 的技术, 可用于研究不同细胞群体或细胞系在药物或基因扰动下的转录组表达情况. 此外, 通过结合细胞哈希 (cell hashing) 技术, MIX-seq 可以进一步探索更为精细的扰动条件, 如治疗后的时间节点或药物剂量等. sci-Plex [65] 将组合索引 (multiplexing) 与高通量单细胞 RNA 测序相结合, 用于研究细胞对多种药物处理或基因干扰的转录组响应. sci-Plex 的核心在于结合单细胞 RNA 测序与细胞核哈希 (nuclear hashing) 的组合条形码策略, 通过组合索引技术, sci-Plex 在同一测序文库中对多个扰动条件下的细胞进行测序, 实现不同实验处理的并行分析. 这种方法大幅度提高了单细胞扰动实验的通量, 同时降低了测序成本和复杂性.

上述技术主要基于单细胞 RNA 测序, 测量扰动后的转录组表达. 目前, 一些新技术进一步将基因编辑与单细胞多组学技术结合, 测量扰动后的多组学特征 (如蛋白质和表观遗传组等). ECCITE-seq [43] 和 Perturb-CITE-seq [19] 实现了对转录本、蛋白质和 sgRNA 的联合检测, 刻画了细胞在受到扰动后基因表达与蛋白质水平的变化. Perturb-ATAC [58] 将 CRISPR 基因干扰与 ATAC-seq (assay for transposase-accessible chromatin using sequencing) 技术结合, 研究基因调控网络和染色质可及性. 在基因扰动后, 利用 ATAC-seq 测量干扰对染色质开放状态的影响, 可提供基因调控和染色质结构变化的高分辨率解析. Spear-ATAC [48] 和 CRISPR-sciATAC [37] 等技术也探索了基因编辑对染色质可及性的影响, 与 Perturb-ATAC 相比, 这些技术提高了实验通量并降低了实验和时间成本.

以上的很多方法都可以进一步用来研究组合扰动的影响. 一些方法通过提高感染复数 (multiplicity of infection) [15, 20], 在一个细胞中引入多个 sgRNA; 另一些则通过使用单个载体来递送多个扰动 [1, 45, 75]. 由于多个 sgRNA 在同一细胞中会增加 Cas9 引发的双链断裂数量, 因此, 一些研究在转录 (CRISPRi)

或 RNA 水平进行扰动. 例如, CaRPool-seq 方法[75] 依赖于 Cas13 在转录水平上进行组合敲低. 由于慢病毒载体在携带 Cas9 sgRNA 数量上存在限制, 一些研究开始使用 Cas12 从单一转录本中处理多个 sgRNA, 从而大幅缩小载体和文库的大小 (参见文献 [13, 80]). 目前, Cas12 也已被应用于 CRISPR 敲除[13]、CRISPR 激活 (CRISPRa)[22] 和 CRISPR 抑制 (CRISPRi)[27] 等多种筛选实验.

随着单细胞扰动技术的不断发展和数据量的快速增长, 已有一些研究尝试整合大规模的单细胞扰动数据集, 并对多种扰动数据进行综合分析. 例如, scPerturb[47] 整合了 44 个公开的单细胞扰动数据集, 涵盖了多种技术和组学, 并分析了不同实验和不同技术之间的数据差异, 包括 sgRNA 的数量、扰动基因的数量、测序深度、细胞总数以及每种扰动的细胞数量等. PerturBase[73] 整理了 46 项公开研究中的 122 个数据集, 包括 115 个单模态数据集和 7 个多模态数据集, 涵盖了 24,254 种基因扰动和 230 种化学扰动, 涉及约 500 万个细胞, 并提供了质量控制、去噪、差异基因表达分析、扰动效应的功能分析以及扰动相似性分析等多种结果. 这些工作为大规模单细胞扰动数据的检索、可视化和整合分析提供了重要的数据支持.

接下来介绍单细胞扰动数据的分析方法. 根据分析目的, 我们将这些分为解析方法和预测方法两类, 并对代表性的方法进行总结. 表 1, 2 和图 2 比较了这些方法的核心模块及输出结果的特点.

表 1　单细胞扰动效应解析方法总结

方法	核心工具	基因层面	基因模块层面	细胞/细胞类型层面	细胞丰度层面	扰动相似性	组合扰动
DESeq2[40]	负二项分布建模	√					
edgeR[54]	负二项分布建模	√					
MAST[18]	双栏模型建模	√					
DESingle[42]	零膨胀的负二项分布模型建模	√					
Milo[11]	k-NN 邻域				√		
MELD[4]	图信号处理 + 顶点频率聚类				√		
PENCIL[53]	带有拒绝的学习策略				√		
scPerturb[47]	E 距离					√	
V2G2P[60]	一致非负正交矩阵分解 + Fisher 精确检验		√				
Interaction manifold[45]	线性模型 + 基因相互作用指标	√					√
MIMOSCA[15]	弹性网络回归	√				√	√
scMAGeCK-LR[78]	岭回归	√					
scMAGeCK-RRA[78]	秩检验	√					
MUSIC[17]	主题模型		√			√	
GSFA[83]	Bayes 因子分析	√	√				
Mixscape[46]	混合判别分析			√			
SCEPTRE[3]	条件随机化检验	√					
CINEMA-OT[16]	独立分量分析 + 最优传输	√		√			√
PopAlign[6]	非负正交矩阵分解 + Gauss 分布建模		√	√	√		

表 2 单细胞扰动预测方法总结

方法	核心工具	基因层面	基因模块层面	细胞/细胞类型层面	实验设计	扰动相似性	组合扰动
D-SPIN [33]	非负正交矩阵分解 + Markov 随机场		√				
CellOracle [35, 36]	整合多模态的自定义 GRN		√				
scGen [39]	变分自编码器 + 潜在空间中的向量运算	√					
CPA [38]	可解释的线性模型 + 对抗自编码器	√					√
Graph VCI [76]	变分 Bayes 因果推断 + 深度图表示学习		√				
GEARS [56]	图神经网络 + 基因间先验知识	√					√
scFoundation [24]	可扩展的 Transformer, 具有非对称编码器 - 解码器	√					√
scGPT [10]	Transformer	√					
Geneformer [69]	Transformer	√	√	√		√	
GeneCompass [79]	Transformer	√					
CellPLM [74]	Transformer	√					
IterPert [28]	预算内的主动学习	√			√		
ALFOI [81]	假设已知因果图的主动学习			√	√		

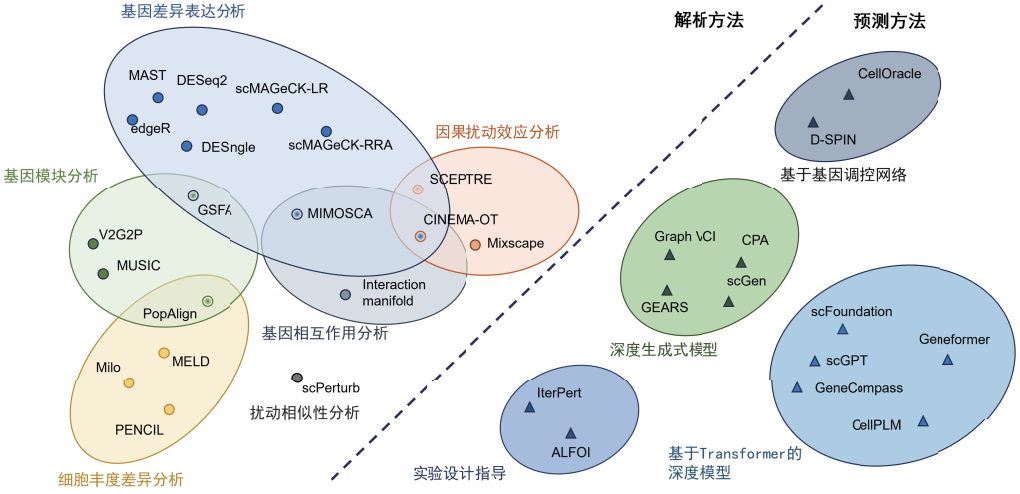

图 2 (网络版彩图) 单细胞扰动数据分析方法总结

3 扰动效应解析方法

量化和解析单细胞扰动数据中的扰动效应是扰动实验要回答的核心问题. 通过比较扰动细胞与对照细胞的组学特征, 利用统计学、机器学习、深度学习的模型和算法, 可以分析基因编辑的下游影响,

理解药物的作用机制. 本节主要聚焦基于单细胞 RNA 测序的转录组扰动数据, 根据算法的特点将扰动效应解析方法分为 5 类: 差异基因表达分析、基因模块分析、细胞丰度差异分析、扰动相似性分析和因果扰动效应分析. 同时, 简要介绍部分代表性方法的流程与原理.

3.1 差异基因表达分析

差异基因表达分析指的是使用统计模型比较不同条件下的基因表达水平, 以识别表达显著变化的基因. 作为一种通用的分析策略, 差异基因表达分析广泛应用于各类转录组数据, 以研究不同生理状态、疾病状态或实验处理条件下的基因表达变化. 在单细胞扰动数据分析中, 差异基因表达分析的主要目标是比较特定扰动 (如基因敲除、药物处理) 与对照条件下的基因表达模式, 以推断扰动对细胞的影响.

差异基因表达分析的关键步骤包括计数数据标准化、选择高变基因、进行适当的统计检验并对多重检验校正, 而后通过可视化和功能富集分析等解读这些差异表达基因的生物学意义. 具体而言, 对于每一个基因, 可以使用均值差异检验、分布差异检验等统计方法 (如 t 检验、Wilcoxon 秩和检验) 来推断该基因在扰动组和对照组的细胞中是否差异表达. 随后, 通过多重检验校正 (如 Benjamini-Hochberg 校正、Bonferroni 校正) 来控制假阳性率. 针对 RNA 测序数据的高噪声、高离散度 (over-dispersion)、非正态性和零膨胀 (zero-inflated) 等特性, 已经开发出多种专门用于 RNA 测序和单细胞 RNA 测序数据的分析方法, 假设它们的模型更符合 RNA 测序数据的特征, 因此能够得到更好的统计功效和假阳性率控制. 这些差异表达分析方法的比较可以参阅最近的文献 [44, 64].

DESeq2 [40] 考虑 RNA 测序计数数据的高离散度, 使用基于负二项分布的广义线性模型 (generalized linear model, GLM) 来进行差异表达分析. DESeq2 假设基因的表达计数服从负二项分布, 并分别通过基因特异的离散度和细胞特异的归一化因子来考虑数据中的高离散度、测序深度及其他技术因素的影响. 在估计基因特异的离散度时, DESeq2 采用经验 Bayes 的方式对这些离散度进行收缩, 将单基因的离散度估计值与整体的趋势进行结合, 以减少因样本或细胞数量有限导致的离散度估计的不稳定性. 差异表达的显著性通过 Wald 检验或似然比检验进行评估. edgeR [54] 也是一种基于负二项分布模型的 RNA 测序数据分析方法, 其模型与 DESeq2 类似, 但在归一化因子的选择、离散度估计方法和假设检验上有所不同. edgeR 采用 TMM (trimmed mean of M-values) 方法进行归一化, 通过去除极端表达值并计算表达值的加权均值来调节样本间的表达差异, 适合处理基因表达水平差异较大的数据集. edgeR 中的离散度估计采用 Cox-Reid 拟似然方法, 该方法同样利用经验 Bayes 收缩, 但基于广义线性模型拟似然的框架. 在假设检验方面, 除了似然比检验, edgeR 还提供了基于拟似然的 QLF 检验 (quasi-likelihood F-tests), 考虑了在基因特异离散度估计时的不确定性, 更好地控制了假阳性率.

MAST [18] 和 DESingle [42] 是专门为单细胞 RNA 测序数据设计的差异表达分析方法, 进一步考虑了表达数据中表达量缺失 (dropout) 的问题. MAST 基于双栏模型 (double-hurdle model) 将表达数据分为离散部分 (是否表达) 和连续部分 (表达量). 对于离散部分, MAST 采用 logistic 回归进行建模; 而对于连续部分, 则采用正态线性模型进行建模. MAST 同样采用经验 Bayes 的方式来收缩基因特异的离散度, 并通过似然比检验推断基因在不同条件下表达的差异性是否显著. DESingle 则基于零膨胀的负二项 (zero-inflated negative binomial, ZINB) 分布对基因的表达计数建模, 并对表达比例差异、表达量差异等不同类型的差异表达基因进行检验.

针对单细胞基因扰动数据, MIMOSCA [15] 和 scMAGeCK [78] 采用带有正则项的线性模型检测差异表达基因. MIMOSCA 将标准化后的基因表达矩阵作为因变量, 将扰动标签及其他混杂因素 (如细

胞状态、测序深度等) 作为自变量, 并通过 ElasticNet 对回归系数进行正则化. 此外, 针对组合扰动数据, MIMOSCA 采用带有交互项的线性模型识别了基因之间的相互作用, 并对相互作用关系相似的基因进行了功能注释. scMAGeCK 包括两个模块: 基于秩检验的 RRA (robust rank aggregation) 模块和基于线性模型的 LR (linear regression) 模块. RRA 模块针对转录组中的每个基因, 计算扰动细胞中该基因表达值在所有细胞中的秩次, 并使用基于次序统计量的假设检验, 评估这些秩次的分位数是否符合均匀分布. 接着通过整合 p 值, 并利用置换检验 (permutation-based test), 确定该基因是否与扰动显著相关. LR 模块与 MIMOSCA 类似, 对所有基因同时进行线性回归, 但区别在于采用 L_2 范数进行正则化, 并通过置换检验筛选出具有显著扰动效应的基因.

3.2 基因模块分析

在差异基因表达分析中, 单细胞扰动数据由于复杂的实验流程、扰动技术因素以及组学测量的影响, 往往表现出稀疏性和高噪声的特征. 此外, 在基因扰动数据中, 扰动类别较多且每种扰动的细胞数量较少. 这些因素都使得对单个基因进行差异表达分析时, 通常会面临统计功效不足或假阳性率较高的问题. 与单基因层面的分析方法不同, MUSIC [17] 和 GSFA [83] 在基因模块 (gene module) 或因子 (factor) 的层面上进行扰动效应的推断. 它们的动机在于基因扰动往往会影响通路中多个基因的表达, 而不是只针对某一特定基因 [80], 通过将功能相似的基因整合成基因模块或因子, 并分析扰动对基因模块或因子的影响, 可以提升统计功效, 并提供清晰的生物学解释.

MUSIC 基于主题模型 (topic model) 对单细胞基因扰动数据进行分析, 其通过数据插补、细胞筛选以及将多个基因整合为主题的方式, 提升了扰动效应推断的统计功效, 同时提供了扰动效应的直接生物学解释. 在预处理阶段, MUSIC 采用数据插补方法 SAVER [29] 来缓解数据的稀疏性, 并设计了细胞筛选流程, 以去除可能存在扰动无效或扰动标签错误的细胞. 在建模阶段, MUSIC 利用主题模型, 将每个细胞的转录组看作一篇 “文章”, 将每个基因的表达量视为一个 “词汇”, 每篇文章都有多个主题的概率分布, 每个主题则有不同的词频概率分布. 利用 Gibbs 采样, 可以获得细胞 - 主题和主题 - 基因的后验分布. 通过比较扰动细胞与对照细胞在主题分布上的差异, MUSIC 能够在主题层面量化扰动效应, 并通过基于主题 - 基因的后验分布进行功能富集分析, 为主题提供功能注释. 此外, MUSIC 还可以在每个主题上或从整体上对扰动效应进行排序、分析扰动之间的相似性.

GSFA 基于 Bayes 因子分析模型进行分析, 其通过因子分解的方式将基因表达矩阵分解为一个潜在因子矩阵和一个因子载荷矩阵, 而后通过线性模型将潜在因子与细胞的扰动标签联系起来. GSFA 分别为因子载荷矩阵和扰动效应矩阵设计了正态混合和针板 (spike and slab) 稀疏先验, 通过 Gibbs 采样的方式得到了后验分布. 此外, GSFA 通过因子载荷将因子层面的扰动效应整合到单个基因水平, 并利用 LFSR 方法 [66] 提供扰动效应显著的基因. 通过对模拟数据和实际数据的分析, GSFA 显示出在扰动效应推断的统计功效方面相较于差异基因表达分析方法的显著优势.

3.3 细胞丰度差异分析

扰动效应可能会改变不同细胞类型或细胞状态下的细胞数量, 因此可以通过比较扰动前后的细胞丰度 (cell abundance) 来量化扰动效应. 最基础的方法是先通过细胞分群、无监督聚类或细胞类型标记等手段确定细胞类型和细胞亚群, 然后在每种细胞类型或亚群中比较扰动前后细胞数量的变化, 并通过卡方检验等研究细胞丰度的变化是否显著 (参见文献 [63]). 但这种方法需要预先指定细胞类型或亚群, 可能引入数据窥探偏差 [2, 9], 且牺牲了单细胞扰动技术的单细胞分辨率. 为了解决这一问题, 研

究者们已经开发出一些基于机器学习或深度学习的方法, 通过比较扰动前后在连续的细胞状态流形或细胞邻域中的细胞丰度, 实现了在更高精度上的细胞丰度差异分析.

Milo[11] 是一种基于 k- 近邻 (k-nearest neighbor, k-NN) 检测细胞丰度差异的方法. 该方法首先将单细胞数据嵌入低维空间, 并通过计算每个细胞的 k- 近邻来构建细胞邻域网络. 每个邻域由具有相似基因表达特征的细胞组成, 代表局部的细胞状态. 而后在构建的 k-NN 图中, 通过比较对照组和扰动组中各邻域的细胞丰度, 计算每个邻域的差异丰度得分 (differential abundance score), 以量化扰动对细胞丰度的影响. 接下来, 使用统计检验评估差异丰度得分的显著性, 并进行多重检验校正, 以确定哪些邻域在不同条件下表现出显著的丰度差异, 从而识别出扰动富集的细胞亚群.

MELD[4] 方法将扰动细胞和对照细胞视为来自底层转录组流形上的概率分布的样本, 结合流形学习和图信号处理的算法在单细胞分辨率上解析扰动效应. 具体而言, MELD 首先采用流形学习技术对单细胞转录组数据进行降维, 将高维数据映射到一个低维的细胞状态流形上. 随后, 利用图信号处理中的图滤波方法, 将每个细胞的扰动标签在流形上进行平滑, 从而计算出每个细胞在不同扰动条件下的概率密度. 接着, 通过比较扰动组与对照组的概率密度, 计算每个细胞属于扰动组或对照组的相对似然值. 这些相对似然值能够揭示细胞状态流形上哪些区域更可能来自扰动组或对照组, 量化扰动带来的丰度差异. 此外, 研究者还提出了一种基于顶点频率聚类 (vertex frequency clustering, VFC) 算法的细胞聚类方法. 不同于基于转录组数据的无监督聚类, 该方法利用相对似然值和细胞扰动标签的频率组成, 对具有相似扰动响应 (扰动富集、不变或对照富集) 的细胞进行聚类, 从而揭示了细胞丰度差异的详细图景.

PENCIL[53] 是一种采用带有拒绝的学习策略 (learning with rejection, LWR) 从单细胞数据中识别与分类或连续表型相关的细胞亚群的方法. 该方法采用监督学习的框架, 设计了扰动标签的预测器与拒绝预测器两个模块, 当拒绝预测器的输出结果大于 0 时进行扰动标签预测, 小于等于 0 时则以一定的代价拒绝预测, 通过综合两个模块的损失函数实现高置信度表型相关的细胞亚群的识别. 该方法在识别扰动相关的细胞亚群时同时通过权重向量进行了基因的筛选, 提供了在扰动和对照细胞亚群划分中重要的基因, 在模拟实验中达到了优于 MELD 和 Milo 的效果.

3.4 扰动相似性分析

从扰动后测量的表型数据中推断扰动之间的相似性是单细胞扰动数据分析的一个重要目标. 通过研究不同基因扰动之间的相似性, 可以识别基因在共同反应途径或网络中的作用. 研究药物扰动之间的相似性则有助于发现具有相似功能的药物群体, 为现有药物在新适应症中的潜在应用提供线索.

以转录组数据为例, 传统方法通过计算扰动细胞在主成分分析等低维空间中的中心点 (centroids) 之间的距离来量化扰动相似性. 然而, 这种方法仅考虑中心点, 忽略了每种扰动的细胞异质性. scPerturb[47] 提出使用 E-distance 来衡量扰动之间的相似性. E-distance 是一种用于衡量两个分布之间距离的统计度量, 较低的 E-distance 表示两种扰动的转录组特征更为相似, 提示它们可能具有相同的下游靶点和通路. 此外, scPerturb 通过计算扰动细胞与对照细胞之间的 E-distance, 研究了不同单细胞测序技术的数据信噪比, 结果显示在不同数据集中, 扰动细胞和对照细胞之间的 E-distance 存在显著差异.

尽管细胞丰度的差异分析、扰动相似性分析能够提供有关扰动效应的重要信息, 但由于单细胞扰动数据的复杂性, 这两类方法在解析扰动效应时仍面临诸多挑战. 一方面, 它们未能明确地将扰动与下游受影响的基因或生物功能直接联系; 另一方面, 对于扰动效应较小、不足以改变细胞状态的数据

集 (如许多基因扰动数据), 难以从细胞丰度角度识别出扰动富集的细胞状态, 也难以从整个转录组层面分析扰动之间的相似性.

3.5 因果扰动效应分析

单细胞扰动数据中存在的多种混杂因素 (如测序深度、实验条件、细胞状态等) 给扰动效应解析带来了挑战. 最近的分析方法从因果推断的角度出发, 有助于实现因果扰动效应的解析.

Mixscape [46] 针对单细胞扰动实验中多种来源的混杂因素, 提出了一套提升扰动效应分析信噪比的流程. 首先, 为了减轻细胞周期阶段 (cell cycle phase)、批次效应 (batch effects) 和内质网应激 (endoplasmic reticulum stress, ER stress) 等差异的影响, Mixscape 将每个细胞的基因表达与该细胞最近的 20 个对照细胞的平均基因表达相减, 从而得到每个细胞的局部基因表达特征 (perturbation signature). 随后, Mixscape 通过混合判别分析 (mixture discriminant analysis) 识别 "扰动逃逸" 的细胞. 对于每个扰动标签, Mixscape 建立了一个两组分 Gauss 混合模型, 两个组分分别代表 "扰动成功" 和 "扰动逃逸", 其中 "扰动逃逸" 组分的分布与对照细胞的分布一致. 最后, 针对 "扰动成功" 的细胞与对照细胞进行比较, 以推断扰动效应. 在对 ECCITE-seq 数据的分析中, Mixscape 为下游分析提供了高信噪比的数据特征.

CINEMA-OT[16] 结合因果学习和最优传输算法, 在单细胞分辨率上推断扰动效应. 首先, CINEMA-OT 通过独立分量分析 (independent component analysis, ICA) 并利用 Chatterjee 系数, 将细胞 - 基因表达矩阵分解为内在状态差异 (混杂因素) 和与扰动相关的因素. 接着, 采用熵正则化的最优传输方法, 生成因果匹配的反事实细胞对 (counterfactual cell pairs). 通过这些反事实细胞对, CINEMA-OT 能够进行单细胞分辨率的扰动效应估计、基于扰动效应的细胞聚类以及协同效应分析等下游应用.

SCEPTRE [3] 提出测序深度等技术因素不仅会影响基因表达的测量, 还会影响扰动标签的检测, 从而产生混杂效应 (confounding effects). 这些混杂因素可能会混淆真正的扰动效应, 或产生假阳性. SCEPTRE 采用条件随机化检验 (conditional randomization test) 的方式去除测序深度带来的混杂效应. 对于每一个扰动 - 基因对, 首先对细胞的扰动标签和测序深度进行 logistic 回归, 计算扰动概率. SCEPTRE 基于该扰动概率对细胞的扰动标签进行重抽样 (resampling), 生成一系列重抽样数据, 而后对原始数据和每个重抽样数据分别拟合负二项分布模型, 计算扰动效应的 z 值 (z-score). 最后, 对重抽样数据的 z 值拟合一个偏斜 t 分布作为零分布 (null distribution), 并将原始数据的 z 值与零分布进行比较, 从而得到修正后的 p 值. 通过条件随机化检验, SCEPTRE 有效修正了测序深度等混杂因素的影响, 在模拟实验和实际数据中表现出更好的假阳性率控制和统计功效.

4 扰动预测方法

尽管目前高通量的扰动技术和检测技术发展迅速, 但在所有可能的细胞类型中通过实验测试所有基因及其组合以及所有化合物及其组合的扰动效应是不切实际的 (参见文献 [38,55]). 因此, 结合目前发展迅速的人工智能和机器学习方法, 对实验未测试的扰动进行预测是很有必要的. 同时, 细胞的调控网络具有明显的结构化特征 (参见文献 [25,49,61]), 对于结构化的系统, 算法便可能通过部分数据学习整个系统的结构并预测其特征. 目前扰动预测方法的目标可以分为两个主要方面: 一是预测扰动效应以探究基因调控机制及基因间的相互作用; 二是预测哪些特定的扰动能够优化迭代模型, 从而有效地指导后续的实验验证 (参见文献 [55]). 这两种方法共同促进了我们对复杂生物系统的理解和控制,

为未来的实验设计提供了理论基础和技术支持. 以下将对这两方面的一些代表性的方法进行总结.

4.1 预测扰动效应的模型

Gavriilidis 等[21] 根据多个指标 (扰动特征、计算特征、数据集特征等) 将预测扰动效应相关的模型分为基因调控网络 (gene regulatory network, GRN) 优先模型 (以 GRN 为先验、主要基于 GRN 构建的方法)、复杂的生成式模型及基础模型 (基于生成式预训练和大量数据的模型) 等. 除此种分类之外, 也可根据模型学习的数据类型将模型分为学习扰动数据的模型和学习生物医学文献的模型 (参见文献 [7,57]), 或可根据解释性将模型划分为融入了 GRN 和先验知识的模型以及难以解释的深度学习模型 (参见文献 [55]). 更多的分类方式和详尽的方法比较可以参考最近的综述 [21,32,55].

(a) 基于基因调控网络的模型. D-SPIN[33] 将细胞建模为一组相互作用的基因表达程序, 并构建一个概率模型 (Markov 随机场或自旋网络) 以推断基因模块与外部扰动之间的调控相互作用. 具体而言, D-SPIN 通过无监督正交非负矩阵分解和功能注释来识别基因表达模块, 接着应用最大似然估计来推断基因模块之间以及模块与施加扰动之间的调控相互作用网络. 相比于大多数非生成式的网络推断方法, D-SPIN 作为一种概率生成式模型可生成扰动下细胞群体的转录状态分布. 将 D-SPIN 应用于人外周血单核细胞 (PBMCs) 数据时, D-SPIN 成功将 500 种免疫调节药物分为 7 类, 并揭示了药物的组合机制. 通过包含数千种扰动条件的单细胞 RNA 测序数据, D-SPIN 提供了一个定量模型, 帮助理解基因调控网络在维持细胞稳态和细胞状态转变中的生物学机理. 该框架还可用于分析不同细胞群体的异质性药物反应, 揭示基因模块的叠加如何诱导产生新的细胞状态.

CellOracle[35,36] 是一种整合多模态扰动数据的 GRN 建模方法, 通过将 scATAC-seq 的启动子和增强子峰与 scRNA-seq 数据结合, 构建了细胞群体特异的 GRN 模型. CellOracle 可以基于模型模拟转录因子的扰动, 研究细胞状态的变化, 并能直观展示这些变化在细胞轨迹图上的映射. 在应用中, 它在小鼠和人类的造血以及斑马鱼的发育过程中验证了因转录因子扰动而导致的细胞表型变化. 此外, CellOracle 提供了对调节细胞状态的转录因子的深入理解, 克服了传统基于深度学习模型的 “黑箱” 问题, 强调了基因调控机制的可解释性, 为细胞状态的机制性分析提供了强有力的工具, 具有广泛的应用潜力.

(b) 复杂生成式模型. scGen[39] 是一种利用变分自编码器在潜在空间中进行向量运算的模型, 以预测细胞对扰动的反应. 该方法假设在潜在低维空间中, 由扰动引起的细胞响应存在线性关系, 可以捕捉高维单细胞数据集中扰动带来的变化. scGen 在预测跨研究和跨物种的扰动影响方面展现出优越的性能, 尤其是在 PBMCs 对 IFN-β 刺激的反应预测中, 其预测性能超越了传统的线性模型、条件变分自编码器和生成对抗网络. scGen 的核心优势在于能够捕捉细胞类型和物种特异性的扰动反应特征, 但其主要针对非基因层面的扰动, 基因相互作用的问题仍未得到充分解决.

CPA[38] (组合扰动自编码器) 通过对抗自编码器框架在潜在空间中寻找药物扰动和细胞状态的嵌入, 以提高基因及药物扰动筛选的预测能力. CPA 结合了线性模型的可解释性与深度学习的灵活性, 能够在未见过的剂量、细胞类型、时间点和物种上进行预测. 其扩展版本 chemCPA[26] 引入了扰动网络, 通过已知的化学描述符对小分子进行编码, 来预测实验未检测化合物的扰动效应. 最新版本 MultiCPA[30] 则结合多模态数据, 利用多模态的数据特征来学习和预测单细胞扰动响应. 在组合扰动预测方面, CPA 能够通过线性模型的叠加填补缺失的药物或基因组合, 促进对组合扰动空间的探索, 从而加速药物和基因研究的实验设计和假设验证.

Graph VCI[76] 结合了深度图表示学习和变分 Bayes 因果推断, 学习未观测到的隐层变量, 并基于

该模型生成因果预测. 该方法利用 GRN 来辅助细胞特异的扰动效应预测, 通过自适应数据更新邻接矩阵, 从而生成与数据相关的精细关系图. 该框架还引入了一种稳健的估计器, 用于有效估计边际扰动效应. 通过广泛的实验, Graph VCI 展现了在个体细胞扰动预测方面, 相较于现有深度学习模型的显著优势.

GEARS [56] (图增强基因激活和抑制模拟器) 结合了图神经网络与基因 - 基因关系知识图谱, 旨在预测细胞对单基因和多基因扰动的转录反应. 该方法利用来自 Perturb-seq 的数据, 通过生物先验和知识图谱获得基因嵌入, 学习基因共表达及基因本体 (gene ontology) 信息. GEARS 通过图神经网络捕捉邻近基因间的信息, 并结合多层感知器进行跨基因的组合, 预测扰动后细胞的基因表达. 在实际应用中, GEARS 展现出优越的性能, 其在组合扰动预测中相较于 CPA 提高了 40% 的精确度, 在预测每种相互作用类型中相互作用最强的 10 个基因对时, 预测准确度是以前方法的 2 倍.

(c) 基于 Transformer 架构的基础模型. scFoundation [24] (又称 xTrimoscFoundation) 是一个拥有 1 亿参数的大规模预训练模型, 基于超过 5,000 万个单细胞 RNA 数据训练而成, 涵盖了大约 20,000 个基因. 该模型通过与已有的预测药物反应的架构 DeepCDR (deep cancer drug response) 以及 SCAD (single cell drug response prediction framework by integrating adversarial domain adaptation) 相组合, 以提升 IC50 值推断和单细胞药物敏感性 (sensitivity) 预测的性能. 其独特的不对称 Transformer 架构能够有效捕捉不同细胞类型及状态下的基因间复杂关系. scFoundation 不仅在基因表达数据增强 (read-depth enhancement)、药物响应预测、单细胞药物响应分类等任务中展示了卓越性能, 还能够进行单细胞基因扰动预测、细胞类型注释和基因模块推断等.

scGPT [10] 是一个基于超过 3,300 万个细胞的单细胞基础模型, 其使用生成预训练 Transformer 技术, 能够在基因维数上实现自注意力机制, 从而编码扰动基因与其他基因间的复杂相互作用. scGPT 通过少量样本学习 (few shot learning) 来学习现有实验数据, 并在未观测的扰动条件下准确预测基因表达. scGPT 不仅展示了在基因表达预测方面的优秀性能, 还能够识别出能够产生特定细胞状态的扰动因子. 通过迁移学习, scGPT 可以优化以执行各种下游任务, 包括细胞类型注释、多批次整合和基因网络推断等, 从而为细胞生物学研究提供有价值的生物学见解.

Geneformer [69] 是一种基于注意力机制和迁移学习技术的深度学习模型. 其将单细胞转录组数据转化为单细胞基因排序数据, 为单细胞转录组提供了非参数化的表示, 并对大规模公共数据库中约 3,000 万个不同组织和细胞类型的人类单细胞转录组数据进行预训练, 以建立编码基因网络动态的基础模型. 此模型在预训练阶段采用完全自监督的方式, 在模型的注意力权重中编码了网络层级结构, 并通过注意力权重获得了对基因网络动态调控的基础性理解. 随后, Geneformer 可以在少量的任务特定数据上进行微调, 适用于包括批次整合、细胞类型注释、基因组元素预测等在内的多种任务. Geneformer 特别适用于数据稀缺的情况, 如罕见疾病研究或难以获取样本的临床研究. 通过在相关任务上的微调, Geneformer 提高了预测准确性, 并成功地在心肌病等 (cardiomyopathy) 疾病中识别出潜在的治疗靶点, 为加速关键调节因子及治疗靶点的发现提供了有力工具.

GeneCompass [79] 是一种创新的知识驱动型跨物种基础模型, 旨在解码生物体内普遍的基因调控机制. 该模型基于超过 1.2 亿个人类与小鼠的单细胞转录组数据进行预训练, 利用 12 层变换器架构和超 1 亿个参数处理复杂的数据结构. 在预训练阶段, GeneCompass 整合了包括启动子序列、基因共表达网络、基因家族信息和转录因子靶基因调控关系在内的 4 种生物学先验知识, 通过自监督的方式加深对基因调控的理解. 此模型不仅在单一物种的多个任务中表现出优于现有最先进模型的性能, 而且还开拓了跨物种生物研究的新方向, 成为理解基因调控机制和药物靶点发现的重要工具.

CellPLM [74] 是一种先进的单细胞预训练模型, 借鉴了大型语言模型的成功经验, 但在设计上针对

单细胞转录组数据的特点进行了优化. 该模型将细胞视作令牌 (token), 而将组织视作句子, 旨在捕捉复杂的细胞间关系. 与传统的基于基因表达序列的方法不同, CellPLM 利用空间分辨的转录组数据来增强对细胞上下文的理解, 并引入 Gauss 混合先验分布以应对数据量有限且噪声大的挑战. 这种方法不仅在多种下游任务中表现出色, 而且推理速度显著提高, 达到了比现有预训练模型快 100 倍的效果. 此外, CellPLM 在空间组学的应用中展现出潜力, 特别是在处理多细胞间关系和空间信息编码方面.

在不同的扰动效应预测模型的研究中, 各模型在其对应的论文中均与其他方法进行了小范围比较. 例如, GEARS 在预测单基因和双基因扰动的前 20 位差异基因时, 其均方误差显著低于基于 GRN 方法的 CellOracle 和复杂生成式模型 CPA. 另外, scGPT 在预测所有基因和差异表达基因的表达变化时, 与实际基因表达变化的 Pearson 相关性也高于 GEARS. 此外, scFoundation 结合了自身模型与 GEARS, 在单基因和双基因扰动的均方误差上也优于原始的 GEARS 模型. 然而, 这些比较都基于特定的数据集和研究范围, 不同研究文章的评估标准存在差异, 可能难以全面衡量各模型在不同任务中的表现.

不同模型对输入数据类型的要求和应用场景也有所不同. 例如, CellOracle 需要同时使用 scRNA-seq 和 scATAC-seq 作为输入数据, 而大多数其他模型则仅需要 scRNA-seq 和扰动数据. GEARS 算法说明中提到, 该方法不适用于跨细胞类型的训练、预测和基于 bulk 数据的训练, 而 CPA 及基于 Transformer 的模型则能够迁移至不同细胞系, chemCPA 可以在 bulk 数据上预训练后进行进一步预测. 此外, 各模型的训练对计算资源的需求也有所不同, 因此在方法选择时需综合考虑数据类型与计算资源的可用性.

4.2 预测扰动用于设计验证实验的方法

IterPert [28] 是一种创新的主动学习方法, 旨在优化 Perturb-seq 实验中扰动的选择, 以克服在有限资源下进行广泛实验设计的挑战. 在实验中, 由于潜在扰动组合数量庞大, 完全测试所有扰动不切实际. 为解决此问题, IterPert 提出了一种迭代方法来设计 Perturb-seq 实验. 它利用现有的模型在每个实验步骤中选择最具信息价值的扰动, 从而有效地缩小搜索范围, 同时保持对未观测到的扰动影响的准确预测. 这种方法特别适用于预算有限的情况, 即所谓的 “有限预算下的主动学习”, 在此情境下, 实验轮次和每次实验的扰动数量都受到了严格的成本和时间限制. IterPert 的有效性通过使用大规模 CRISPRi Perturb-seq 数据集构建的计算基准得到了验证. 结果显示, 与其他方法相比, IterPert 能够在使用更少的实验次数的情况下达到相似的预测精度. 这表明 IterPert 在指导 Perturb-seq 实验设计方面具有显著优势, 特别是在资源受限的情况下.

ALFOI [81] (active learning for optimal intervention) 是一种因果主动学习策略, 目的是在高维扰动空间中高效识别出能够最优化扰动后分布均值与预期均值差异的扰动策略. 此方法具体可分为连续的两步: (1) 使用现有从不同扰动获得的样本更新模型信念 (model belief); (2) 通过构建和优化一个获取函数选择下一个扰动以获取样本, 从而确保所选样本能提供关于期望结果的最有价值的信息. 这一获取函数以封闭形式进行计算, 便于快速优化. 相比于目前其他利用因果结构的 Bayes 优化方法, ALFOI 在以下两方面进行了拓展: (1) 优化整个分布均值, 而不是单个目标节点; (2) 考虑连续取值的扰动, 而不是离散或有限扰动. ALFOI 适用于多种场景, 如基因组学中的细胞重编程、机械系统的反馈控制、气候变化研究中的干预设计等, 尤其在处理连续值扰动时表现出色, 如药物剂量的优化调整. 实验证明, 在处理模拟数据及单细胞转录组数据时, ALFOI 的表现优于现有的标准方法, 能够在较少但精细挑选的样本上实现更优的扰动设计.

5 结论与展望

本综述系统总结了当前单细胞扰动技术的发展及其产生的数据, 包括 Perturb-seq, CROP-seq 和 MIX-seq 等. 这些技术结合了 CRISPR 或药物筛选与单细胞测序, 为揭示基因调控、细胞状态转变和药物响应等复杂生物学过程提供了前所未有的解析能力. 我们进一步介绍了单细胞扰动数据的分析方法, 如 MUSIC, GSFA 和 scGen 等. 这些方法各具特色, 从基于传统统计推断和潜在空间学习的模型, 到结合机器学习、深度学习和大语言模型的先进方法, 为单细胞扰动效应的解析和预测提供了多样化的工具.

单细胞扰动技术和数据分析方法在未来仍具有广阔的发展前景. 首先, 技术的提升将着重于克服现有方法的局限, 特别是在动态和时空分辨率方面的改进. 现有的技术主要集中于静态数据的收集, 难以准确捕捉细胞在动态环境中的时空变化. 此外, 现有的数据集缺乏大规模、多模态、纵向和空间分辨率等更高复杂性的特征, 这限制了我们对细胞状态和药物作用机制的深入理解. 未来的单细胞扰动技术将朝着多样化和高复杂性方向发展, 例如以 Perturb-Map [14] 为代表的创新技术, 将扰动与空间转录组学相结合, 有望在原位条件下更全面地揭示细胞对扰动的响应. 在下游应用方面, 大规模项目如 The Lifetime Consortia [51] 将提供大量与疾病相关的组学数据资源, 推动单细胞扰动技术在疾病研究、药物开发和筛选中的应用. 此外, 通过构建一个大型的细胞扰动图谱 (perturbation cell atlas) [55], 可以与当前的人类细胞图谱 (human cell atlas, HCA) [52] 进行结合, 进一步促进我们对细胞生物学的深刻理解, 并为疾病治疗和药物开发提供更为精准的支持.

在扰动效应解析方面, 面对单细胞扰动数据的高噪声及各类混杂因素, 如何从数据中进行因果学习以识别真正的扰动效应仍是值得深入研究的问题. 针对扰动效应的预测, 许多新兴的机器学习、深度学习方法以及大型基础模型已展现出巨大的潜力, 但模型的可解释性和实际生物学应用中的有效性仍需进一步验证和优化. 因此, 提升模型的可解释性将是未来研究的重点. 通过大型基础模型、因果表示学习及与生物学先验知识的融合, 研究人员有望更准确地预测扰动对不同细胞的影响, 并验证相关科学问题.

总之, 单细胞扰动技术和数据分析方法的持续进步, 将为我们理解细胞在扰动状态下的复杂行为提供更为丰富的信息. 随着技术的不断成熟和数据资源的日益丰富, 这一领域有望在精准医学、药物研发及其他生物医学应用中发挥关键作用, 为实现个性化治疗奠定坚实的基础.

致谢 图 1 通过 BioRender.com 制作.

参考文献

1 Adamson B, Norman T M, Jost M, et al. A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. Cell, 2016, 167: 1867–1882
2 Assmann S F, Pocock S J, Enos L E, et al. Subgroup analysis and other (mis)uses of baseline data in clinical trials. Lancet, 2000, 355: 1064–1069
3 Barry T, Wang X, Morris J A, et al. SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. Genome Biol, 2021, 22: 344
4 Burkhardt D B, Stanley J S III, Tong A, et al. Quantifying the effect of experimental perturbations at single-cell resolution. Nat Biotechnol, 2021, 39: 619–629
5 Carter P J, Rajpal A. Designing antibodies as therapeutics. Cell, 2022, 185: 2789–2805
6 Chen S, Rivaud P, Park J H, et al. Dissecting heterogeneous cell populations across drug and disease conditions with PopAlign. Proc Natl Acad Sci USA, 2020, 117: 28784–28794
7 Chen Y Q, Zou J. GenePT: A simple but effective foundation model for genes and cells built from ChatGPT. bioRxiv:2023.10.16.562533, 2024

8 Cheng J, Lin G, Wang T, et al. Massively parallel CRISPR-based genetic perturbation screening at single-cell resolution. Adv Sci, 2023, 10: 2204484
9 Cook D I, Gebski V J, Keech A C. Subgroup analysis in clinical trials. Med J Australia, 2004, 180: 289–291
10 Cui H, Wang C, Maan H, et al. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods, 2024, 21: 1470–1480
11 Dann E, Henderson N C, Teichmann S A, et al. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol, 2022, 40: 245–253
12 Datlinger P, Rendeiro A F, Schmidl C, et al. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods, 2017, 14: 297–301
13 DeWeirdt P C, Sanson K R, Sangree A K, et al. Optimization of AsCas12a for combinatorial genetic screens in human cells. Nat Biotechnol, 2021, 39: 94–104
14 Dhainaut M, Rose S A, Akturk G, et al. Spatial CRISPR genomics identifies regulators of the tumor microenvironment. Cell, 2022, 185: 1223–1239
15 Dixit A, Parnas O, Li B, et al. Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. Cell, 2016, 167: 1853–1866
16 Dong M, Wang B, Wei J, et al. Causal identification of single-cell experimental perturbation effects with CINEMA-OT. Nat Methods, 2023, 20: 1769–1779
17 Duan B, Zhou C, Zhu C, et al. Model-based understanding of single-cell CRISPR screening. Nature Commun, 2019, 10: 2233
18 Finak G, McDavid A, Yajima M, et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol, 2015, 16: 278
19 Frangieh C J, Melms J C, Thakore P I, et al. Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. Nat Genet, 2021, 53: 332–341
20 Gasperini M, Hill A J, McFaline-Figueroa J L, et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. Cell, 2019, 176: 377–390
21 Gavriilidis G I, Vasileiou V, Orfanou A, et al. A mini-review on perturbation modelling across single-cell omic modalities. Comput Struct Biotechnol J, 2024, 23: 1886–1896
22 Griffith A L, Zheng F, McGee A V, et al. Optimization of Cas12a for multiplexed genome-scale transcriptional activation. Cell Genomics, 2023, 3: 100387
23 Haber A L, Biton M, Rogel N, et al. A single-cell survey of the small intestinal epithelium. Nature, 2017, 551: 333–339
24 Hao M, Gong J, Zeng X, et al. Large-scale foundation model on single-cell transcriptomics. Nat Methods, 2024, 21: 1481–1491
25 Heimberg G, Bhatnagar R, El-Samad H, et al. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. Cell Syst, 2016, 2: 239–250
26 Hetzel L, Boehm S, Kilbertus N, et al. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In: Advances in Neural Information Processing Systems. Red Hook: Curran Associates, 2022, 26711–26722
27 Hsiung C C, Wilson C M, Sambold N A, et al. Engineered CRISPR-Cas12a for higher-order combinatorial chromatin perturbations. Nat Biotechnol, 2024, 43: 369–383
28 Huang K X, Lopez R, Hütter J C, et al. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. In: Research in Computational Molecular Biology. Lecture Notes in Computer Science, vol. 14758. Berlin: Springer, 2024, 17–37
29 Huang M, Wang J, Torre E, et al. SAVER: Gene expression recovery for single-cell RNA sequencing. Nat Methods, 2018, 15: 539–542
30 Inecik K, Uhlmann A, Lotfollahi M, et al. MultiCPA: Multimodal compositional perturbation autoencoder. bioRxiv: 2022.07.08.499049, 2022
31 Jaitin D A, Weiner A, Yofe I, et al. Dissecting immune circuits by linking CRISPR-pooled screens with single-cell RNA-Seq. Cell, 2016, 167: 1883–1896
32 Ji Y, Lotfollahi M, Wolf F A, et al. Machine learning for perturbational single-cell omics. Cell Syst, 2021, 12: 522–537
33 Jiang J L, Chen S S, Tsou T, et al. D-SPIN constructs gene regulatory network models from multiplexed scRNA-seq data revealing organizing principles of cellular perturbation response. bioRxiv:2023.04.19.537364, 2023
34 Jiang L, Dalgarno C, Papalexi E, et al. Systematic reconstruction of molecular pathway signatures using scalable single-cell perturbation screens. Nat Cell Biol, 2025, 27: 505–517
35 Kamimoto K, Adil M T, Jindal K, et al. Gene regulatory network reconfiguration in direct lineage reprogramming. Stem Cell Rep, 2023, 18: 97–112
36 Kamimoto K, Stringa B, Hoffmann C M, et al. Dissecting cell identity via network inference and in silico gene

perturbation. Nature, 2023, 614: 742–751
37 Liscovitch-Brauer N, Montalbano A, Deng J, et al. Profiling the genetic determinants of chromatin accessibility with scalable single-cell CRISPR screens. Nat Biotechnol, 2021, 39: 1270–1277
38 Lotfollahi M, Susmelj A K, De Donno C, et al. Predicting cellular responses to complex perturbations in high-throughput screens. Mol Syst Biol, 2023, 19: e11517
39 Lotfollahi M, Wolf F A, Theis F J. scGen predicts single-cell perturbation responses. Nat Methods, 2019, 16: 715–721
40 Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol, 2014, 15: 550
41 McFarland J M, Paolella B R, Warren A, et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. Nat Commun, 2020, 11: 4296
42 Miao Z, Deng K, Wang X, et al. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics, 2018, 34: 3223–3224
43 Mimitou E P, Cheng A, Montalbano A, et al. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. Nat Methods, 2019, 16: 409–412
44 Nguyen H C T, Baik B, Yoon S, et al. Benchmarking integration of single-cell differential expression. Nat Commun, 2023, 14: 1570
45 Norman T M, Horlbeck M A, Replogle J M, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science, 2019, 365: 786–793
46 Papalexi E, Mimitou E P, Butler A W, et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. Nat Genet, 2021, 53: 322–331
47 Peidli S, Green T D, Shen C, et al. scPerturb: Harmonized single-cell perturbation data. Nat Methods, 2024, 21: 531–540
48 Pierce S E, Granja J M, Greenleaf W J. High-throughput single-cell chromatin accessibility CRISPR screens enable unbiased identification of regulatory networks in cancer. Nat Commun, 2021, 12: 2969
49 Prieto C, Risueño A, Fontanillo C, et al. Human gene coexpression landscape: Confident network derived from tissue transcriptomic profiles. Plos One, 2008, 3: e3911
50 Rafelski S M, Theriot J A. Establishing a conceptual framework for holistic cell states and state transitions. Cell, 2024, 187: 2633–2651
51 Rajewsky N, Almouzni G, Gorski S A, et al. LifeTime and improving European healthcare through cell-based interceptive medicine. Nature, 2020, 587: 377–386
52 Regev A, Teichmann S A, Lander E S, et al. The human cell atlas. Elife, 2017, 6: e27041
53 Ren T, Chen C, Danilov A V, et al. Supervised learning of high-confidence phenotypic subpopulations from single-cell data. Nat Mach Intell, 2023, 5: 528–541
54 Robinson M D, McCarthy D J, Smyth G K. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 2010, 26: 139–140
55 Rood J E, Hupalowska A, Regev A. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. Cell, 2024, 187: 4520–4545
56 Roohani Y, Huang K, Leskovec J. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. Nat Biotechnol, 2023, 42: 927–935
57 Roohani Y, Vora J, Huang Q, et al. BioDiscoveryAgent: An AI agent for designing genetic perturbation experiments. arXiv:2405.17631, 2024
58 Rubin A J, Parker K R, Satpathy A T, et al. Coupled single-cell CRISPR screening and epigenomic profiling reveals causal gene regulatory networks. Cell, 2019, 176: 361–376
59 Schenone M, Dančík V, Wagner B K, et al. Target identification and mechanism of action in chemical biology and drug discovery. Nat Chem Biol, 2013, 9: 232–240
60 Schnitzler G R, Kang H, Fang S, et al. Convergence of coronary artery disease genes onto endothelial cell programs. Nature, 2024, 626: 799–807
61 Segal E, Shapira M, Regev A, et al. Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet, 2003, 34: 166–176
62 Seo S, Lee T K, Kim M H, et al. Prediction of side effects using comprehensive similarity measures. Biomed Res Int, 2020, 2020: 1357630
63 Shifrut E, Carnevale J, Tobin V, et al. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. Cell, 2018, 175: 1958–1971
64 Squair J W, Gautier M, Kathe C, et al. Confronting false discoveries in single-cell differential expression. Nat Commun, 2021, 12: 5692
65 Srivatsan S R, McFaline-Figueroa J L, Ramani V, et al. Massively multiplex chemical transcriptomics at single-cell

resolution. Science, 2020, 367: 45–51
66 Stephens M. False discovery rates: A new deal. Biostatistics, 2017, 18: 275–294
67 Szałata A, Hrovatin K, Becker S, et al. Transformers in single-cell omics: A review and new perspectives. Nat Methods, 2024, 21: 1430–1443
68 Tanay A, Regev A. Scaling single-cell genomics from phenomenology to mechanism. Nature, 2017, 541: 331–338
69 Theodoris C V, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. Nature, 2023, 618: 616–624
70 Thorman A W, Reigle J, Chutipongtanate S, et al. Accelerating drug discovery and repurposing by combining transcriptional signature connectivity with docking. 2024, 10: eadj3010
71 Tian R L, Abarientos A, Hong J S, et al. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. Nat Neurosci, 2021, 24: 1020–1034
72 Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol, 2016, 34: 1145–1160
73 Wei Z, Si D, Duan B, et al. PerturBase: A comprehensive database for single-cell perturbation data analysis and visualization. Nucleic Acids Res, 2024, 53: D1099–D1111
74 Wen H, Tang W, Dai X, et al. CellPLM: Pre-training of cell language model beyond single cells. bioRxiv: 2023.10.03.560734, 2023
75 Wessels H H, Méndez-Mancilla A, Hao Y, et al. Efficient combinatorial targeting of RNA transcripts in single cells with Cas13 RNA Perturb-seq. Nat Methods, 2023, 20: 86–94
76 Wu Y L, Barton R A, Wang Z C, et al. Predicting cellular responses with variational causal inference and refined relational information. arXiv:2210.00116, 2022
77 Xie S, Duan J, Li B, et al. Multiplexed engineering and analysis of combinatorial enhancer activity in single cells. Mol Cell, 2017, 66: 285–299
78 Yang L, Zhu Y, Yu H, et al. scMAGeCK links genotypes with multiple phenotypes in single-cell CRISPR screens. Genome Biol, 2020, 21: 19
79 Yang X, Liu G, Feng G, et al. GeneCompass: Deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. Cell Res, 2024, 34: 830–845
80 Yao D, Binan L, Bezney J, et al. Scalable genetic screening for regulatory circuits using compressed Perturb-seq. Nat Biotechnol, 2024, 42: 1282–1295
81 Zhang J Q, Cammarata L, Squires C, et al. Active learning for optimal intervention design in causal models. Nat Mach Intell, 2023, 5: 1066–1075
82 Zheng X H, Wu B L, Liu Y J, et al. Massively parallel *in vivo* Perturb-seq reveals cell-type-specific transcriptional networks in cortical development. Cell, 2024, 187: 3236–3248
83 Zhou Y, Luo K, Liang L, et al. A new Bayesian factor analysis method improves detection of genes and biological processes affected by perturbations in single-cell CRISPR screening. Nat Methods, 2023, 20: 1693–1703

A review on single-cell perturbation data analysis

Zichu Fu, Ming Yang & Lin Hou

Abstract Single-cell perturbation technologies combine external perturbations, such as gene editing or drug screening, with single-cell sequencing, providing detailed molecular profiles of cellular responses at single-cell resolution. Computational analysis of these data aids in uncovering gene regulatory networks and drug mechanisms, as well as predicting the potential effects of unobserved gene or drug combinations. This review summarizes key single-cell perturbation technologies and highlights a growing range of methods for interpreting and predicting perturbation effects, extending from classical statistical models to machine learning and deep learning approaches. We also discuss the challenges and prospects, including perturbation atlases, multi-omics, spatial data, and causal learning.

Keywords perturbation, single-cell sequencing, machine learning, deep learning

MSC(2020) 62P10, 92B20, 68T09

doi: 10.1360/SSM-2024-0315