npj | systems biology and applications
www.nature.com/npjsba

评论文章 开放式

Check for updates

基因调控网络重建：利用单细胞多组学数据的力量

Daniel Kim[1,2,3,5] , Andy Tran[1,3,4,5] , Hani Jieun Kim[2,3] , Yingxin Lin[1,3,4] , Jean Yee Hwa Yang[1,3,4] ✉ and Pengyi Yang[1,2,3,4] ✉

推断基因调控网络（GRN）是生物学中的一项基本挑战，旨在揭示基因及其调控因子之间的复杂关系。破译这些网络对于理解驱动许多细胞过程和疾病的潜在调控串扰起着至关重要的作用。测序技术的最新进展推动了利用匹配的单细胞多组数据的最先进 GRN 推断方法的发展。通过采用不同的数学和统计方法，这些方法旨在重建更全面、更精确的基因调控网络。在这篇综述中，我们将简要介绍 GRN 推断方法常用的统计和方法论基础。然后，我们比较和对比了最新的单细胞匹配多组学数据 GRN 推断方法，并讨论了它们的假设、局限性和机遇。最后，我们讨论了这一快速发展的领域所面临的挑战和有望取得进一步进展的未来方向。

npj 系统生物学与应用 (2023) 9:51 ;https://doi.org/10.1038/s41540-023-00312-6

引言

基因的转录调控是所有重要细胞过程的基础，由许多分子调控因子错综复杂的相互作用来协调。[1].基因调控的前沿是转录因子（TFs），它们与 DNA 中称为顺式调控元件（CREs）的特定区域（如启动子和增强子）相互作用。[2,3].TFs、CREs 和基因之间的相互作用共同构成了基因调控网络（GRNs）。[4]并在各种疾病的发生和发展中发挥重要作用。[5].随着高通量组学技术的发展，对涉及基因调控的许多分子特征进行剖析已成为可能。然而，这些网络的重建面临着巨大的挑战，需要开发强大而高效的计算工具来揭示基因调控网 络 的调控相互作用。

最早的计算 GRN 推断方法是利用微阵列和 RNA 测序（RNA-seq）技术的数据开发的，这些技 术 可 定量测量整个细胞群的 RNA 表达（图 1）。[6].这些方法利用互信息和相关性等关联测量方法识别共表达基因，从 而 确定潜在的调控关系。[7,8].然而，这些方法无法纳入驱动基因调控的表观遗传变化信息，从而限制了它们评估调控结合位点（包括 TFs 结合位点）可 及 性 的 能力。从大容量转录组学扩展到大容量多组学（图 1）测序技术（如 ATAC-seq）后，这些局限性得到了缓解；[9]Hi-C，这是一种测量全基因组染色质构象的技术，可捕捉结构变化和染色质相互作用；[10]以及 ChIP-seq，该技术可捕获全基因组蛋白质与 DNA 的相互作用，包括增强子和启动子的 TF 结合位点。[11].然而尽管批量测序技 术 能够揭示机理，更可靠地捕捉调控关系，但它们缺乏捕捉细胞类型和/或特定状态信息的能力。

单细胞全息技术的出现彻底改变了我们以单细胞分辨率揭示细胞异质性的能力（图 1）。[12].由单细胞 RNA-seq (scRNA-seq)[13]、单细胞 ATAC-seq (scATAC-seq)[9]和单细胞 ChIP-seq 等技术产生的数据。[14]和单细胞 ChIP-seq (scChIP-seq)[15]这些方法现在可以在细胞类型、细胞状态和单细胞水平上推断调控因子与其靶基因之间的调控关系。[16–18].此外，单细胞组学技术已从分析单一模式（如 scRNA-seq、scATAC-seq）发展到以单细胞分辨率捕获多种模式（即 "单细胞多组学"）。[19].特别是，一系列新型测序平台有能力同时分析单细胞内的 RNA 和 CRE 可及性，如 SHARE-seq 和 10x Multiome。[20,21].因此，这些技术推动了新的 GRN 推断方法的发展，利用这些数据可以进一步全面地再现细胞类型和细胞状态水平的调控网络。[22,23].

然而，浏览众多 GRN 推断方法并了解它们如何推断调控连接 是一项具有挑战性的任务，对于可能没有定量背景的研究人员来说尤其如此。此外，可用的 GRN 推断方法数量众多，很难确定最适合特定研究的方法。

感兴趣的问题。为此，我们对最新开发的用于配对 scRNA-seq 和 scATAC-seq 数据的 GRN 推断方 法 进行了分类，回顾了 GRN 推断的方法论基础，旨在为研究人员和方法开发人员提供帮助。我们首先简要介绍 GRN 推断方法的历史及其从体细胞测序到

[1]澳大利亚新南威尔士州坎珀当悉尼大学数学与统计学院。[2]澳大利亚新南威尔士州坎珀当，悉尼大学，儿童医学研究所，计算系统生物学组。[3]澳大利亚新南威尔士州坎珀当悉尼大学悉尼精密数据科学中心。[4]悉尼大学查尔斯-帕金斯中心
澳大利亚新南威尔士州坎珀当悉尼。[5]这些作者的贡献相同：Daniel Kim、Andy Tran。电子邮件：jean.yang@sydney.edu.au; pengyi.yang@sydney.edu.au

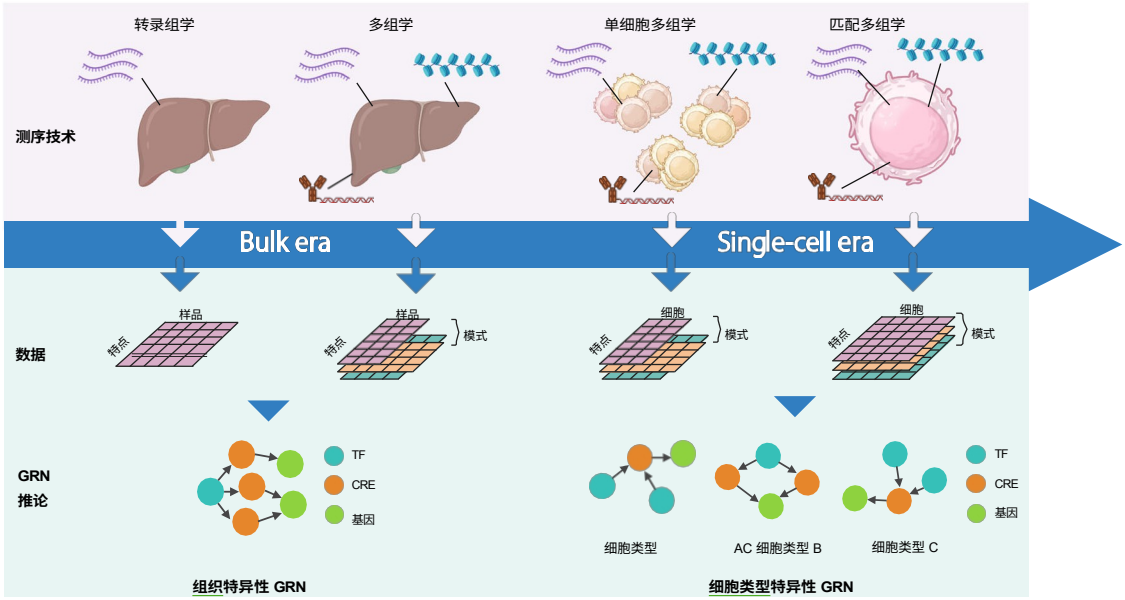

Fig. 1 Schematic illustration of the parallel development and evolution of GRN inference and sequencing technologies. Initially, bulk sequencing technologies provided insights into regulatory interactions at the tissue level but were limited in capturing cellular heterogeneity. The emergence of single-cell technologies revolutionized the field, enabling the inference and reconstruction of cell type-specific gene regulatory networks. The advancements in sequencing technologies now allows for the multi-omic profiling of cells, offering a remarkable opportunity to precisely capture and integrate diverse molecular signals within the same cell, as shown in the cell furthest to the right of Fig. 1. Importantly, each sequencing technology possesses its own unique data structure and characteristics. For example, data of unmatched modalities do not share identical dimensions, as the cells and features, including their respective numbers, differ between each modality. Consequently, integration methods are required to map cells and features into a common space prior to GRN inference. In contrast, matched multi-modal data do not require data integration as the different modalities are captured within the same cell, which minimizes noise and thus improves the quality and accuracy of GRN inference. As a result of the developments in sequencing technologies and data structures, more accurate and comprehensive regulatory networks may be reconstructed. It is important to note that not all single-cell GRN methods reconstruct cell type or state-specific regulatory networks but instead take advantage of additional omics layers to better represent regulatory network architectures.

技术，包括常用的 GRN 推断方法的基本理论基础。如需更全面的概述，建议读者阅读之前广泛介绍早期 GRN 推断方法的评论文章[5,24–26].因此，我们将详细评述最近使用单细胞配对多组数据重建 GRN 的方法，包括其优势和潜在的局 限 性 。最后，我们讨论了 GRN 推断方法目前面临的挑战和潜在的发展方向，希望能对该领域未来的方法发展有所启发。

Grn 推论的方法论基础

GRN 推断依赖于统计和算法原理，以揭示基因及其调控因子之间的调控联系。通过利用相关性、回归、概率模型、动态系统和深度学习等各种技术（图 2），研究人员可以有效地建模和推断生物系统底层的调控架构。在此，我们将简要讨论常用的统计方法以及当前配对多组学数据 GRN 推断方法的基本假设。

基于相关性的方法

重建基因组网络最常用的方法之一是 "因关联而有罪"（guilty by association）的概念。换句话说，，共同表达的基因被假定为功能相关或共同调控。例如，TF 及其假定靶基因的共同表达可能表明两者之间存在调控关系。同样，CREs 及其目的基因也可以被认为是通过将 CRE 的可及性与推定靶基因的表达水平相关联来确定。常用的关联测量方法包括参数皮尔逊相关性和非参数斯皮尔曼相关性，它们可分别捕捉线性和非线性关联（图 2）。线性相关能有效检测 TF 表达或 CRE 可及性增加导致基因表达成比例变化的关系。然而，非线性相关可以捕捉到更复杂的关系，从而更好地再现 TF、CRE 和基因之间的调控相互作用。[27].其他方法包括互信息，这是一种基于信息论的非参数方法，用于测量两个变量之间的依赖关系。[8].

虽然相关性分析可为潜在的调控关系提供有价值的见解，但必须注意的是，仅靠相关性分析有明显的局限性。例如，如果两个 TF 的表达水平相关，相关性不能确定哪个是调控因子和靶标，也不能排除它们受第三个 TF 调控的可能性。此外，相关性测量难以区分直接或间接关系，包括可能存在混杂因素时。不过，纳 入 ATAC-seq 等其他模式的信息有 可能缓解这些局限性，因为它们提供了额外的证据，证明调控因子与下游靶基因之间存在定向关系，即 TF 必须与染色质的可访问区域结合才能调控其靶基因。

回归模型

回归法提供了一种捕捉响应变量与多个预测变量之间关系的方

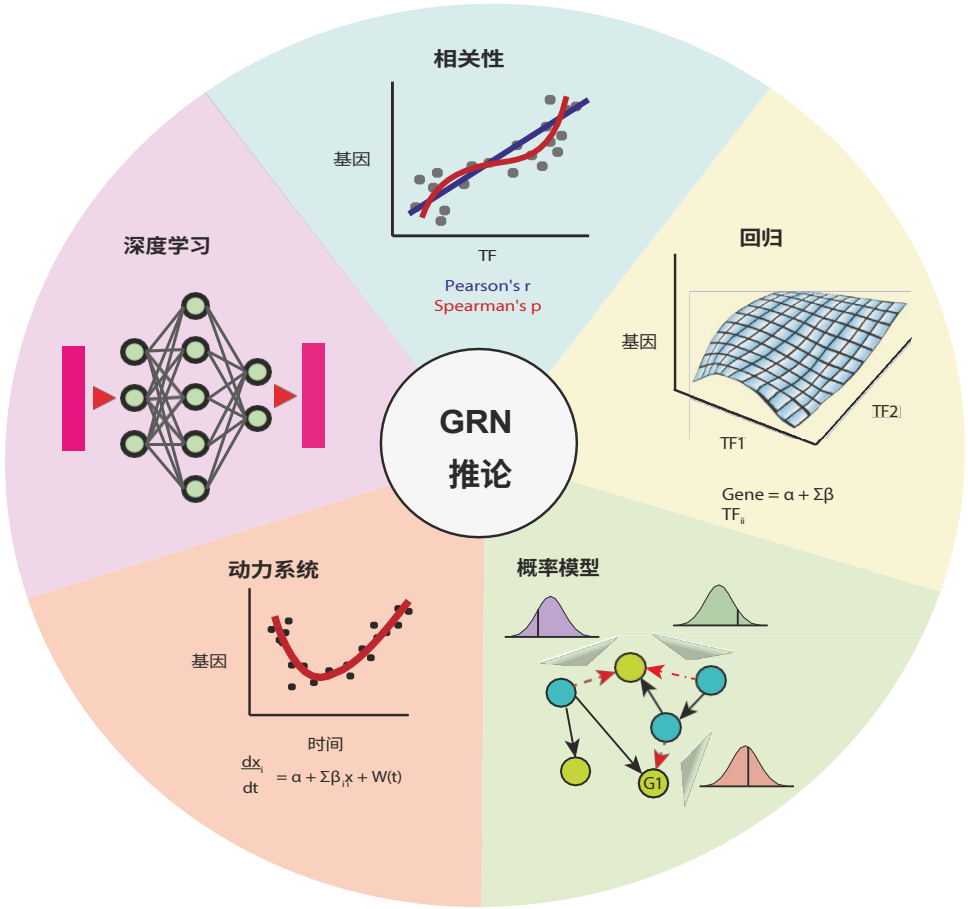

图 2 配对单细胞多组学 GRN 推断方法的主要类别。基于相关性的方法旨在找出变化相似的成对变量（即 TF 表达、基因表达或 CRE 可及性）。基于回归的方法根据多个预测变量（即 TF 表达和/或 CRE 可及性）建立基因表达模型。概率模型旨在确定基因最可能的调节因子。基于动态系统的方法根据生物因素（如 TF 表达、细胞周期阶段、一般随机性）对基因表达的变化进行建模。基于深度学习的方法利用神经网络推断 TF、CRE、基因和细胞之间的复杂关系。

在 GRN 推论中，响应变量可以是一个基因的表达量，并分别与多个 TFs 和 CREs 的 表达量或可及性进行回归（图 2）。通过明确估计每个预测因子对响应（如基因表达）的影响，回归模型的系数（如 TFs 或 CREs）可解释为关联的强度，而系数的符号可用于推断调控相互作用的方向。

在使用普通最小二乘回归法推断 GRN 的情况下，数据可能包含数千个 TF 或 CRE，这取决于从目标基因转录起始位点搜索的距离。[28].重要的是，加入大量预测因子往往会导致过拟合，即模型变得过于复杂，泛化效果不佳。此外，如果存在相关的预测因子， 回归模型就会变得不稳定，而在生物环境中，由于 TFs 可以相互调控，这种情况很可能发生。为了解决这些问题，更现代的惩罚回归方法（如 LASSO）根据系数的绝对大小引入了一个额外的惩罚项，有效地将选定的系数缩减为零，从而降低了最终估计调控网络的复杂性。此外，非参数方法（如基于树的回归）并不假定数据中存在任何固定结构，但其可解释性较差，构建时的计算量也更大。

概率模型

用于 GRN 推断的概率模型一般采用图形模型的形式，该模型捕捉变量（如 TF 及其靶基因）之间的依赖关系。这些方法一般旨在模拟每个 TF 与其假定的靶基因之间是否存在调控关系和/或调控关系的强度，而这种强度是通过找出能解释给定训练数据的最可能的关系来估算的。通过这些概率度量，可以在下游分析之前对调控相互作用进行筛选和优先排序，从而进行更有针对性的研究。然而，这些方法通常假设基因表达遵循特定的分布，如高斯分布，但这一假设可能并不适合所有基因的表达。[29].

动力系统

回归法和基于概率的方法直接根据预测变量对响应变量进行建模，而基于动力系统的方法则试图对随时间演变的系统行为进行建模。就 GRN 推断而言，人们可能有兴趣估计基因的表达与各种因素的关系，如 TF 的调控效应、基础转录和随时间变化的一般随机性（图 2）。这些效应可以作为微分方程中的参数建

模，而微分方程可以通过数据或文献进行估计[30].

与之前讨论的方法相比，动态系统模型具有明显的优势，因为它们能捕捉到影响基因表达及其随机性的各种因素。估算出的模型是可解释的，每个参数都对应一个特定的属性。然而，大型网络的复杂性和对先验特定领域知识的依赖性会降低这些模型的可扩展性，并容易产生发表偏差。[31,32].

深度学习模型

深度学习模型是一类机器学习技术，近年来在包括生物信息学在内的众多学科中备受关注。[33].这些模型以人工神经网络为基础，可用于执行各种任务（图 2）。[33–35].例如，多层感知器可以解决回归式问题以估计函数，而自动编码器则可用于降维。特别是，自编码器可以有多种类型的输入，并学习它们之间的共同连接，代表潜在的调控关系。[36].

然而，深度学习方法的灵活性是以为代价的，因为它们通常需要非常大的训练数据集。

最小建模假设。此外，构建的模型通常包含大量参数，需要大量计算资源才能估算出来。与传统统计模型相比，深度学习方法的可解释性通常也较差，因为拟合系数通常没有明确的解释。[37].不过，最近的一系列方法，如显著性，旨在通过识别整体模型中的重要特征来纠正这一问题，这些特征可用来识别候选的 TF 调节器。[38].

批量 Omics 时代的基因推断

大容量转录组学

微阵列和 RNA 测序（RNA-seq）等高通量图谱分析方法是最早捕捉样本全局转录组图谱的实验方法之一。[39].为此，人们开发了计算方法，通过分析成千上万个基因的表达模式，揭示转录因子与其靶基因之间的潜在调控联系。[40].著名的例子包括 ARCANE、CLR 和 MRNet，它们利用互信息等关联指标来量化转录因子与其靶基因之间的关系。[41–43].然而，这些方法的一个关键制约因素在于它们的成对关联计算，无法将基因表达模拟为多个调控因子的函数。基于回归的方法（如 GENIE3）将基因表达作为多个调控因子的函数来建模，从而解决了这一限制，这种方法可以更准确地建模调控因子和目标基因之间的调控关系。[44,45].尽管如此，这些方法的一个重要局限是仅依赖于转录组学数据，从而忽略了已知在基因调控中起关键作用的表观遗传修饰。

批量多组学

基因调控和转录过程有许多分子机制和参与者，如表观遗传修饰因子，它们通过复杂的相互作用来调控基因的表达。这些分子调控因子在启动、促进、增强和调控基因转录方面发挥着重要作用。因此，要构建更全面的基因转录网络，必须纳入更多的调控因子和 DNA 元素（如增强子和沉默子）以及结构信息，包括

染色质构象。例如，ATAC-seq 可用于生成更全面的 GRN，如 GRaNIE、PECA 和 TimeReg 所用的那样[46–48].DISTILLER 和 ChIP-Array 2 等方法整合了 RNA 和 ChIP-seq 数据，以确定目标基因的 TF 和调控序列。[46,47,49–51].Hi-C 还可用于捕捉 DNA 的构象，并与 ATAC-seq 和 RNA-seq 数据整合，构建多组学 GRN。[11,52].总之，整合各种多组学数据集并使用统计模型有可能加深我们对基因调控的理解，并揭示 TF 与其靶基因在不同生物环境中的动态相互作用。[30,53,54].

尽管批量转录组学和批量多组学 GRN 推断方法各有优势，但也有共同的局限性。任何仅基于批量数据的分析都很难推断出细胞类型的特异性信息，因为omics 图谱是整个细胞群体的平均值，从而消除了细胞异质性的任何信号。[55].然而，众所周知， 各种疾病，如糖尿病和癌症，全部或部分是由特定的细胞类型群驱动的。[56,57].

单细胞时代的基因推断

单细胞分子生物学

单细胞整体组学技术的诞生缓解了大量整体组学技 术 在推断基因组网络方面的许多局限性。这些技术提供了对不同组织细胞和分子组成的详细了解，超越了批量测序方法的能力[13,58–60].通过 scRNA 测序，转录组学率先进入单细胞水平。许多流行的 GRN 方法都是为利用 scRNA-seq 数据而设计的，包括基于回归（SCENIC、scTenifoldNet）、动力系统（SCODE）和信息论（PIDC）的方法。[22,61–63].

如今，测序技术可通过 scATAC-seq、sciHi-C 和 scChIP-seq 对其他模式进行量化，从 而 有助于全面捕捉细胞内分子间的动态变化。[9,15,59].DeepTFni 等方法已被开发出来，可独立利用这些附加模式，为 GRN 推断提供另一种方法。[64].其他方法旨在结合多种模式的信息。例如，CellOracle、MICA 和 IReNA 分两个阶段分别使用 scRNA-seq 和 scATAC-seq，其中包括过滤推测的调控联系，然后构建最终的 GRN，反之亦然。[65–67].或者，也可以从不同的模式中构建单独的 GRN，然后合并生成一个综合 GRN。[68].

目前已开发出一系列其他方法，用于整合来自不同细胞的多组学数据，同时学习不同模式之间的共享关系，以重建调控网络。其中包括 DC3、scREG 和 scAI，它们使用矩阵因 式 分 解 技术将不匹配的多组学数据投射到低维表示中，从而将它们整合在一起。[69–71].同样，GLUE 和 scTIE 通过将不同的模式投射到低维嵌入中来整合多组学数据，但它们使用的是自动编码器，这是一种基于深度学习的技术，可以从数据中推断出复杂的结构。[72,73].一旦学习到能捕捉 omics 层之间共享模式的低维表示，这些方法就能利用映射提取多组学特征，从而推断出相互作用（如 CRE 与基因之间的相互作用），这些相互作用可用于重建 GRN。这些方法也可应用于匹配的 scRNA-seq 和 scATAC-seq 数据，将它们视为独立的细胞群。不过，由于这些方法的主要目的不是用于 GRN 推断，因此我们不在本 文 中对其 进 行 评述。

实现匹配的单细胞多组学

从批量 RNA 到批量多组学的发展涉及到更多模式的开发和整合，多模式单细胞 omics 技术带来了新一轮的技术浪潮，这些技术可以对同一细胞内的不同模式（通常称为匹配或配对数据）进行剖析。[74].这些技术包括 SNARE-seq，它可以联合剖析转录组和染色质的可及性。[75]CITE-seq，一种捕获转录组和细胞表面蛋白标记的方法[21]配对标记，这是一种同时分析组蛋白修饰和转录组的 高通量方法[76]和 ASAP-seq，以单细胞分辨率捕获转录组 、染色质景观和蛋白质标记表达。[77].重要的是，测序技术的这些进步为利用多模态数据中蕴含的信息 提供了机会，而整合不匹配的多组学数据时可能无法获得这些信息。不过，目前已开发出一系列计算技术来匹配不同模式的单细胞，或对缺失的模式进行估算，从而提高多模式单细胞数据的可用性和可访问性。[78,79].

最新的 GRN 推理方法旨在利用这些优势
新数据来建立更全面的基因调控模型，从而推断出更稳健、更复杂的调控网络。然而，它们的方法和复杂程度各不相同，而且并非所有单细胞多组学基因调控网络推断方法都能重建细胞类型或特定状态的调控网络。因此，可能很难理解它们之间的差异以及在不同情况下的适用性。在此，我们将最新的配对多组学数据 GRN 方法分为五大类（相关、回归、概率模型、动态系统和深度学习），并讨论它们的共同点和不同点。必须承认的是，这些分类并不能完全概括每种方法所采用的全部统计和方法框架，因为许多方法结合了多种技术来重建 GRN。不过，通过简化分类，我们希望让读者对指导这些方法的基本原理有一个广泛而易于理解的认识。图 3 列出了这些方法。我们希望这份全面的概述能帮助研究人员了解当前 GRN 推理方法的发展情况，并促进在应用这些方法时做出明智的决策。

基于相关性的方法

这些方法利用相关性推断成对调控元件之间的潜在调控关系，如 CRE vs 基因或 TF vs CRE（图 4）。这些方法只考虑与推定靶基因的 TSS 之间用户指定距离内的 CRE，TF-CRE 连接的推断通常包括 TF 主题富集分析（图 4）。虽然基于相关性的方法乍看起来很相似，但它们在相关性指标的选择和实施方面存在一些关键差异。例 如 ，STREAM 和 scMEGA 使用皮尔逊相关性来捕捉线性关系，而 FigR 和 TRIPOD 则使用斯皮尔曼相关性来捕捉非线性关系。[33,59,80,81].

FigR 和 STREAM 的目的是识别调控模块，这些模块捕捉细胞类型或状态中的关键过程。简而言之，FigR 筛 选 具有调控染色质域（DORC）的基因，DORC 的定义是具有用户定义数量的显著相关 CRE 的基因。因此，FigR 能生成专门由 DORC 组成的 GRN。同样，STREAM 构建的网络模块由共同表达的基因和可共同访问的 CREs 组成。然后， ，通过主题富集分析确定这些模块最可能的调控 TF。

另外，scMEGA 和 TRIPOD 的目标是识别构成整个 GRN 的各个调控环节。

TF-CRE-基因的可及性和基因表达，包括TF表达和基 因 表达，来选择候选的TF-基因调控配对。然而，TRIPOD 的目标是找到 TF-CRE 基因的调控三元组。三元组是通过计算基因表达与 TF 表达和 CRE 可及性的相关性来确定的，同时将已确定的 CRE 基因和 TF 基因关联作为另一个组成部分的 条件。更确切地说，CRE-基因关系是以 TF 表达为条件的，方法是匹配 TF 表达值最接近的细胞对，并利用 CRE 可及性和基因表达的差异进行相关性分析。因此，检测到的 CRE 基因联系不会受到 TF 表达的干扰。同样，TF-基因关系以 CRE 可及性为条件，以考虑到不同的 CRE 可及性会影响 TF 结合能力，从而调节基因表达。[52].

基于回归的方法

考虑到基因可能有多个 TF 调控因子，反之亦然，DIRECT-NET、SCENIC +、Pando、scRE- MOTE 和 RENIN 利用回归法将基因表达模拟为多个调控因子的函数。这些方法又可分为参数回归法（Pando、scREMOTE 和 RENIN）和非参数回归法，如基于树的回归法（DIRECT- NET 和 SCENIC +）（ 图 5）。

一种方法是普通最小二乘法回归，其最简单的形式是假设基因与其调控因子之间存在线性关系。Pando 和 scREMOTE 将基因表达建模为 TF 表达和 CRE 可及性的线性函数[23,82].Pando 通过直接将基 因 表 达 量 与 CRE 可及性和 TF 表达量的乘积进行回归，来估计每个 TF 对基因的调控效应；而 scREMOTE 则将调控潜能作为回归中的一个权重，该权重根据 TF 主题富集、CRE 可及性和染色质构象进行估计。在此基础上，RENIN 使用了两个带有自适应弹性网估计器的模型，这种正则化技术会对大系数进行惩罚，从而使调控网络更加稀疏，假阳性结果更少。[83].第一个模型捕捉 CRE 可及性与基因表达之间的关系，以识别可能调控目标基因的 CRE。第二个模型是 TF 表达和基因表达的模型，它结合了第一个模型的结果来识别 TF 与基因的联系。在所有情况下，线性模型的推断系数都可以解释为 TF 对目标基因的调控效应，从而构成 GRN。重要的是，Pando、scREMOTE 和 RENIN 的一个明显缺点是它们仅限于识别 TF 和 CRE 等调控因子与其靶基 因 之间的线性关系。

DIRECT-NET 和 SCENIC+ 可以通过使用一种称为梯度树提升的基于树的回归算法来捕捉非线性关系，从 而 缓解这一局限性。[17,22,52].DIRECT-NET 提供了一种有价值的功能，因为它可以计算每个 CRE 在预测基因表达方面的可及性的重要性，然后在推断 TF 基因链接之前将它们标记为高、中或低置信度的 CRE。这样就能更好地控制，因为只有高置信度的 CRE 才会被保留用于进一步的下游分析。虽然 DIRECT-NET 和 SCENIC

+ 使用 TF 主题富集建立 TF 基因对，SCENIC
+ 使用的是内部生成的图案汇编，其中包含 30,000 多个独特的位置权重矩阵，每个 TF 平均有 5 个指定图案。在预测 TF 结合位点时，这可能比将其归纳为共识序列（如典型的主题富集分析中使用的序列）更有优势，因为它可以捕捉到更广泛的 TF。

概率模型

迄今为止讨论的方法都是将目标基因与其他基因分开考虑，而概率模型则不同，它可以模拟基因之间的协方差。为此，单细胞多任务

图 3 本综述所包括的当前配对多组学数据 GRN 推断方法摘要。Prog.，实现该方法的程序；Cell type，该方法是否产生细胞类型特异性 GRN；Metacell，该方法是否将单细胞聚合成元细胞（多个相似细胞的平均表达谱）；TF-CRE，用于推断 TF-CRE 链接的主要方法；CRE-Gene，用于推断 CRE 基因链接的主要方法；TF-gene，用于推断 TF 基因链接的主要方法。输入：RNA，该方法是否只与 scRNA-seq 数据兼容；ATAC，该方法是否只与 scATAC-seq 数据兼容。Prob.，概率模型；D.S，动力系统；D.L，深度学习；diff.eq.，微分方程。不适用单元格表示特定方法未完成识别 TF-CRE、CRE-基因或 TF-基因联系的相应步骤。有关每种方法的更多详情，请参阅*单细胞时代的GRN推断*一节。

网络推断（scMTNI），旨在通过采用贝叶斯框架，在估计细胞类型特异性调控网络时纳入调控关系的先验知识，重建细胞类型或条件特异性 GRN（图 6）。[84].

scMTNI 使用细胞系树将以下假设纳入其中相关细胞类型应具有相似的 GRN，以及 相应的 scATAC-seq 数据，以优先选择在目标基因的可访问启动子区域中具有图案的 TF 调控因子。TF 基因网络是通过概率图形模型推断的，将每个基因的表达视为随机变量，并以一组 TF 调控因子为条件。该模型的估算方法是，从一个空的 TF 基因调控列表开始，反复添加最有可能解释目标基因表达的调控连接。通过两个可调参数，用户可以限制推断出的 GRN 中的边缘数量，并对基因启动子区域中的 TF 矩阵的重要性进行加权。最终输出是用户提供的细胞系树中每种细胞类型的 GRN。重要的是

请注意，scMTNI 假设基因表达遵循高斯分布，这可能并不代表生物现实[85].此外，基于贝叶斯的方法的输出可能对先验的选择很敏感，从而可能限制推断出的 GRN 的稳健性。[86].

基于动力系统的方法

目前讨论的 GRN 推理方法一般假定相关细胞群足够均匀，任何变化都是由噪声引起的。然而，单个细胞之间的差异可能具有生物学意义，并受到细胞周期及其环境的影响。将这些因素纳入 GRN 推断过程具有明显的优势，因为它考虑到了基因调控和环境相互作用的动态性质。在这种情况下，Dictys 可通过伪时间分析捕捉轨迹上的静态和时间分辨 GRN（ 图 7）。[80].

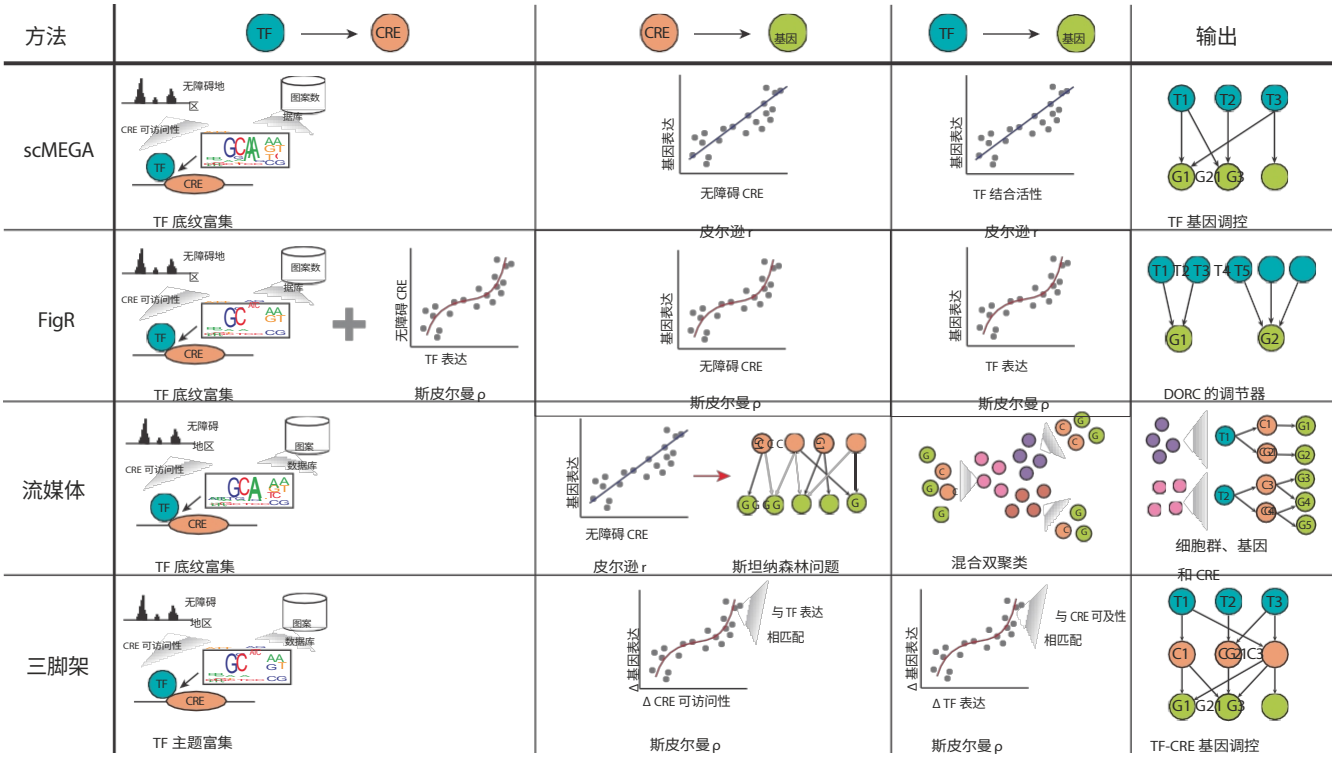

图 4 基于相关性的方法示意图。使用皮尔逊相关性的方法（scMEGA 和 STREAM）仅限于检测线性调控关系，而使用斯皮尔曼相关性的方法（STREAM 和 TRIPOD）则既能捕捉线性关系，也能捕捉非线性关系。scMEGA、FigR 和 STREAM 直接使用细胞的基因表达、CRE 可及性和 TF 表达测量值进行相关性分析，而 TRIPOD 则使用与其他成分匹配的细胞之间的差异。DORCs 调控染色质域，T 转录因子，C 顺式调控元件，G 基因。

与迄今为止讨论过的方法不同，Dictys 同时以 TF 基底和基元为目标建立 TF-CRE 链接，其中 TF 基底较小，因此不易被检测出假阳性。[80]其中 TF 脚印较小，因此不容易被检测为假阳性。[80].然后，每个基因的潜在调控因子会被筛选为能与附近 CRE 结合的 TF。然后用经验线性模型将 TF 与推定靶基因之间的关系建模为随机微分方程，最终拟合系数代表 TF 对其靶基因的调控效应。值得注意的是，Dictys 可以恢复差异调控（logFC）和差异表达（CPM）。使用差异调控可以帮助模拟 TF 与其目标基因之间调控活性的变化，这种变化并不完全依赖于基因表达水平。因此，由于 Dictys 模拟的是随时间变化的表达，它可能更适合研究 GRN 内的差异调控变化，特别是在细胞分化等连续过程中。此外，由于 Dictys 采用核平滑法构建调控模型，因此它对因观测数据较少而导致的高变异性具有很强的鲁棒性。不过，值得注意的是，与线性回归一样，Dictys 将总的调控效应估计为单个 TF 表达式的线性组合，这可能过度简化了真正的生物学关系，而这种关系往往更为复杂。[81].

基于深度学习的方法

深度学习模型因其学习复杂非线性模式的能力而备受关注，并在生物医学成像、蛋白质结构预测和蛋白质功能预测等多个领域取得了巨大成功。[33,34].最近的一些研究利用深度学习模型来利用最近获得的单细胞配对多组学数据来推断调控网络，包括 DeepMAPS、MTLRank 和 LINGER（图 8）。[38,87,88].

与其他经审查的 GRN 推断方法不同，DeepMAPS 和 MTLRank 纳入了 RNA 速度（定义为剪接和未剪接信使 RNA 的比率），该速度可估算特定基因在测序时的基因表达变化率。[89].TF 对其靶基因的调控影响很少是瞬时的，它涉及一连串的调控事件（共调控蛋白的招募和染色质重塑），最终导致基因表达的变化。因此，将 RNA 速度作为基因表达随时间变化的代理变量，可以在估计和确定 TFs 对其目标基因的调控作用时提供更准确的近似值。

DeepMAPS 对每个基因的调控潜力进行了估算。通过汇总 CRE 的可及性及其与基因转录起始位点的邻近程度，可以得出每个细胞中每个基因的调控潜力和 RNA 速度。然后将调控潜力和 RNA 速度汇总为基因活动矩阵，以捕捉每 个 细胞中每个基因的动态特性。然后使用图自动编码器学习基因和细胞的低维嵌入，并将 具有相似基因活动的细胞和基因分组。然后为每个细胞群建立基因之间的调控联系。与迄今为止讨论过的大多数方法一样，DeepMAPS 利用 CRE 中的 TF 底物富集来推断这些簇的调控 TF。相比之下，MTLRank 通过 ChIP-seq 和 scATAC-seq 数据计算 TF 活性得分，以估计细胞中每个 TF 和基因之间的调控效应。然后将 TF 活性与 TF 表达相结合，利用多层神经网络预测 RNA 速度。然后，MTLRank 根据 TF 对其推定靶基因 RNA 速度的影响对 TF 进行排序，从而推断出调控关系，进而重建 GRN。

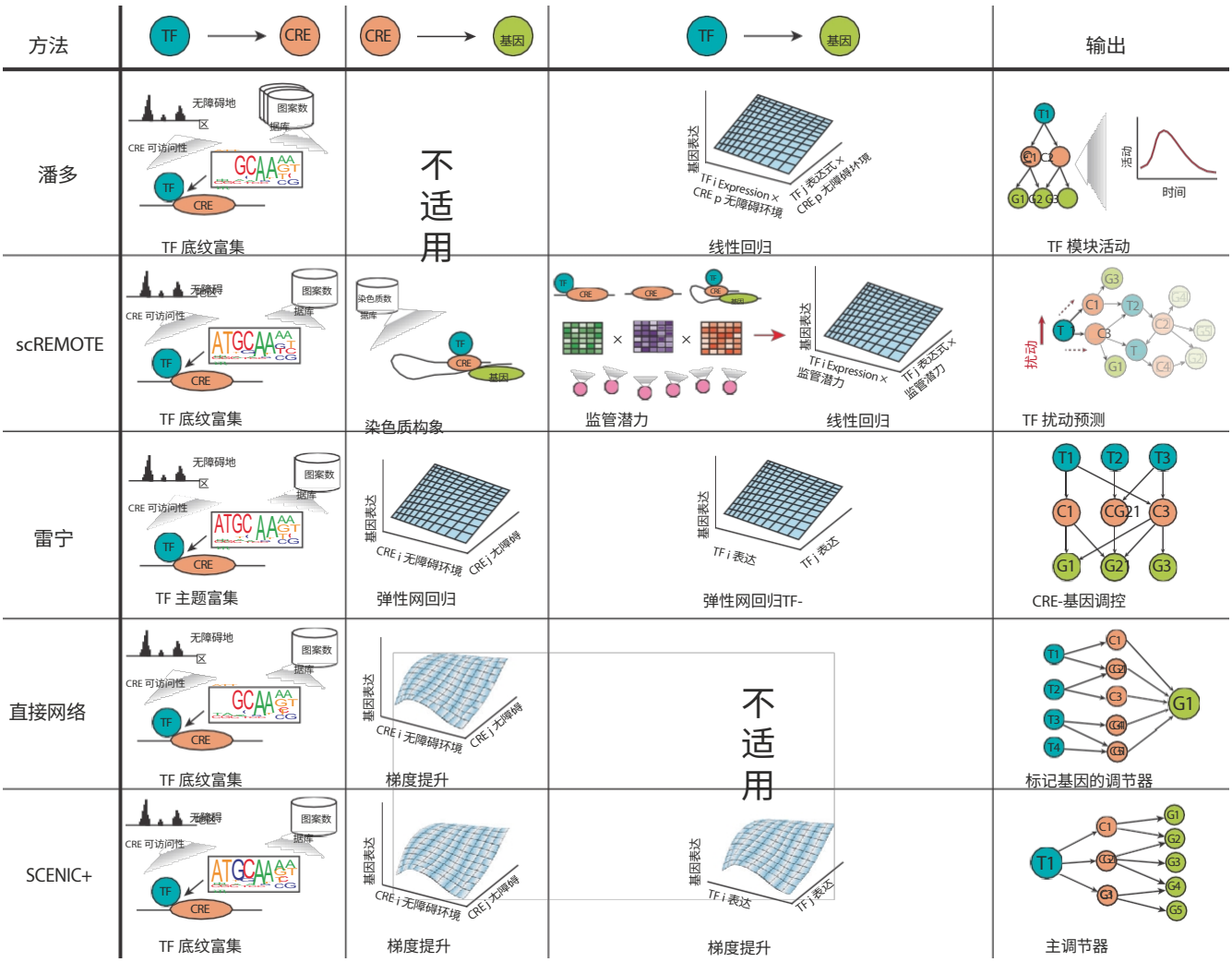

图 5 基于回归的方法示意图。Pando、scREMOTE 和 RENIN 只限于推断线性调控关系。DIRECT-NET 和 SCENIC+ 使用梯度提升法来捕捉非线性关系。T 转录因子，C 顺式调控元件，G 基因。

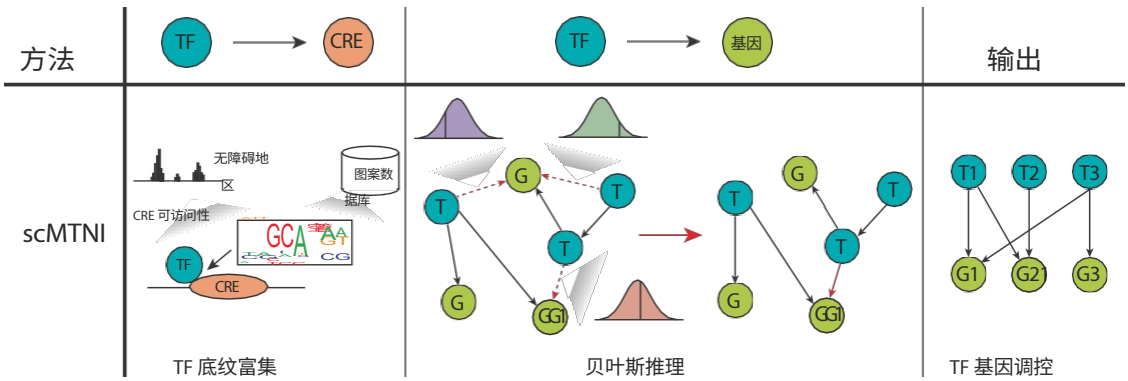

图 6 scMTNI 的示意图。scMTNI 利用贝叶斯推理来确定最可能的 TF 基因规程，从而构成 GRN。在 TF 到基因的步骤中，实线箭头表示推断出的边缘，红色虚线箭头表示候选边缘。最有可能的候选边缘被添加到推断边缘列表中。T 转录因子，C 顺式调节元件，G 基因。

或者，LINGER 直接使用 TF 表达和 CRE 可及性，结合 TF 主题富集，利用多层神经网络预测基因表达。LINGER 首先在批量数据上对网络进行训练，这样做的好处是可以充分利用跨越多种背景的图集级数据知识。训练的结果

然后利用匹配的 scRNA-seq 和 scATAC-seq 数据对网络进行完善。与 MTLRank 相似，TF 和 CRE 的 调控重要性也是通过它们对推测的靶基因表达水平的影响来估算的。此外，TF 与 CRE 之间的联系还可以通过它们在第一序列中的权重和第二序列中的权重之间的相关性来推断。

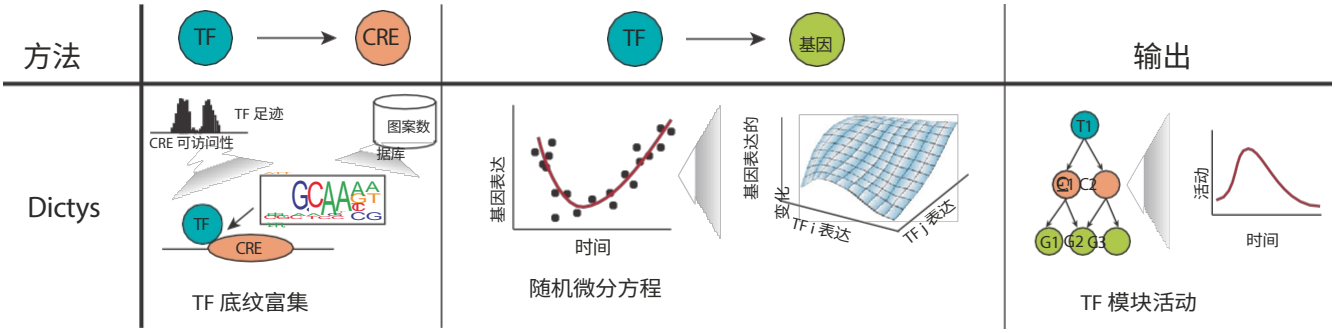

图 7 Dictys 示意图。Dictys 利用随机微分方程将基因表达模拟为多种调控因子和 因子的函数。输出结果可解释为随时间变化的调控活动。T 转录因子，C 顺式调节元件，G 基因。

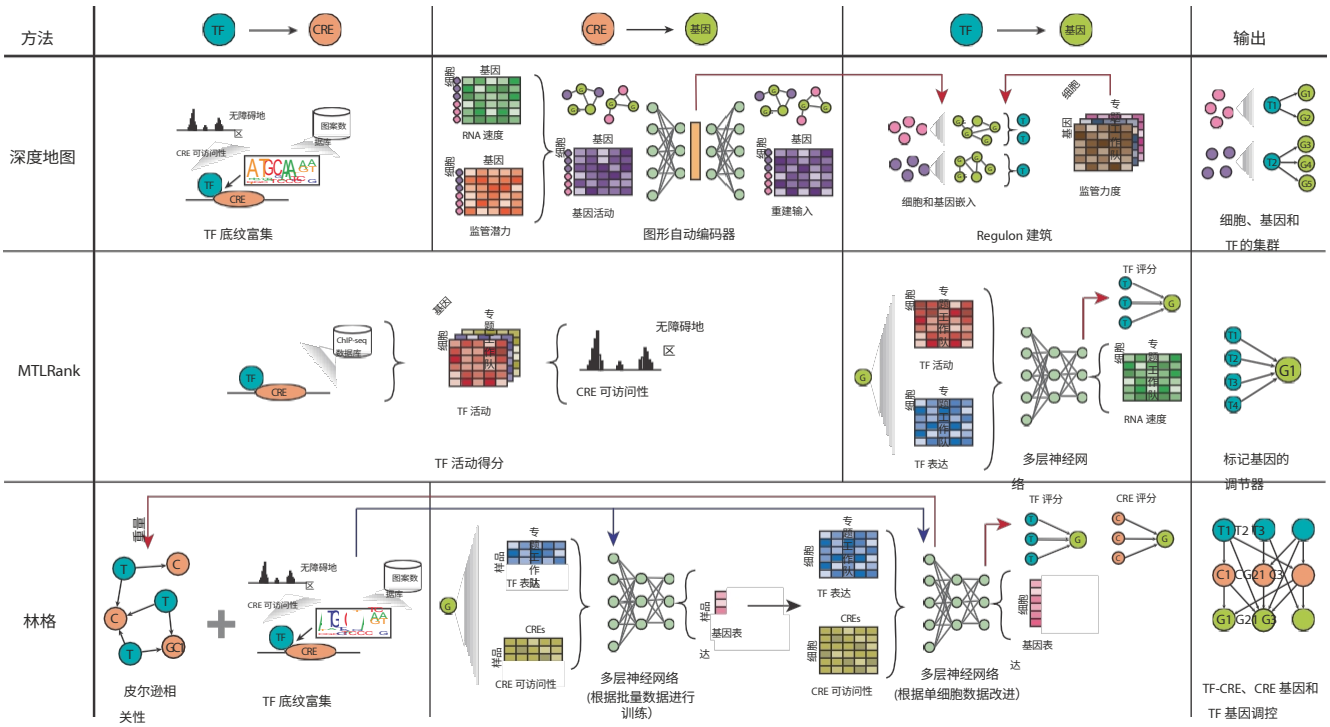

图 8 基于深度学习的方法示意图。DeepMAPS 和 MTLRank 都利用 RNA 速度来推断调控关系。DeepMAPS 为细胞群识别这些网络，而 MTLRank 则专门重建细胞类型标记基因的调控连接。不过，LINGER 使用 TF 表达和 CRE 可及性来预测基因表达，先在批量数据上进行训练，然后在单细胞数据上进行改进。LINGER 将 TF 主题富集纳入神经网络结构，并利用学习到的权重推断 TF-CRE 联系。T 转录因子，C 顺式调节元件，G 基因。

这使得 LINGER 能够构建所有 TF-CRE、CRE-基因和 TF-基因链接，从而重建 GRN。

挑战与机遇

尽管 GRN 推理算法取得了重大进展，但仍存在一些关键限制。在此，我们将讨论这些挑战和未来改进的潜在机会。

数据稀疏性

与批量数据相比，单细胞数据通常具有明显的稀疏性和噪声，这可能会影响鲁棒 GRN 的构建。[90].例如，据估计，批量数据中零的比例约为 10%-40%，而单细胞数据中零的比例可高达 90%。[91] 而单细胞数据中的零比例可高达 90% 。[92].单细胞数据的稀疏性可部分归因于技术原因，例如低效的库制备和

序列扩增[93].此外，单细胞技术旨在捕获单个细胞的表达谱，而单个细胞往往表现出许多基因的低表达水平，导致捕获的 RNA 转录本数量有限。与此相反，批量测序技术汇总了许多细胞的分子表达谱，可以捕获更多的计数，但 会损失细胞类型层面的异质性信息。重要的是，单细胞数据中存在高比例的零会导致对基因表达相关性的估计出现偏差和不稳定，从而使 GRN 的准确推断变得更加复杂。[94].许多 GRN 推断方法旨在通过将多个相似细胞聚合成元细胞（多个相似细胞的平均表达谱）来解决这些问题。然而，这可能会导致相关性膨胀，从而有可能推断出错误的调控关系[95,96].其他策略包括估算，即使用各种方法估算缺失值，包括概率模型和潜在空间嵌入。然而，大多数现有的估算方法

主要是为 scRNA-seq 数据的估算而 设计的，对其他数据模式的可用选项有限[97].不过，随着测序技术的不断进步，测序深度的提 高 ， 我们预计这一领域会有重大发展。此外，许多专门用于处理稀疏数据的统计和生物信息学方法已经出现，这表明了在 GRN 推断中管理数据稀疏性的方法论进步[38,98].

确定因果关系

GRN 推断的另一个重大挑战是建立调控因子与其目标基因之间的因果关系。大多数方法都是通过某种关联度来推断调控关系，如相关性[99].同样，回归和概率方法也能模拟变量间关联的强度和方向。[100].然而，由于可能存在混杂因素，仅靠这些指标和模型不足以建立因果调控关系。然而，整合捕捉基因调控不同方面（如染色质可及性和构象）的多种恳求模式，可以为真正的调控联系提供进一步的证据。例如，TF 结合位点与其靶基因之间存在染色质环，这 表 明 TF 可与靶基因的调控区域（如启动子或增强子区域）进行物理结合，从而表明两者之间存在调控关系。[54].此外，扰动或时间序列实验等实验方法提供了一种更直接的方法，通过扰动调控因子并观察其各自的靶基因表达水平随时间的变化来推断调控联系。[101,102].例如，如果扰动 TF 会导致抑制或激活其靶基因的表达水平，那么 TF 与靶基因之间就更有可能存在调控关系。在同一细胞内捕捉这些信号凸显了匹配多组学数据的优势，因为不同模式之间的关系来自相同的生物背景，从而提高了调控联系的 质量和准确性。

验证

鉴于重建 GRNs 的目的是再现感兴趣的生物过程，因此对 GRNs 进行验证是一项关键的公开挑战。因此，GRN 验证需要彻底调查重建的 GRN 与 "地面实况 "之间的一致性。要做到这一点，从湿实验室实验（如功能扰动实验）中推断出的基本真实调控网络至关重要。[103].功能缺失和功能增益实验通常通过观察调控因子表达水平的 变化是否会导致其假定靶基因的激活或抑制，从而更有把握地建立调控联系。[101,104].CRISPR- cas9 技术的出现使这些调控相互作用的高通量筛选成为可能，大大提高了扰动实验的效率和产出。[105].利用 CRISPRi 增强子平铺筛选，还可以针对增强子等非编码区，量化 CREs 的 变化可能对下游靶基因产生的影响，从而为建立 CREs 和靶基因之间的真正调控联系提供了一种手段。[106].值得注意的是，实验验证可能既费钱又费时，对于匹配的剖析技术来说尤其如此。然而，ISSAAC-seq 等测序技术的进步为单细胞模式的联合剖析提供了更经济实惠的选择，并为更好地利 用 匹配剖析技术铺平了道路。[107].因此，我们预计，随着测序成本因效率和灵敏度的提高而降低，重构 GRN 的实验验证将变得更加普遍。

制定基准

同样，有必要对 GRN 推断方法进行验证和基准测试，以改善目前的局限性。GRN推断方法在其重建的调控网络中表现出相当大的多样性，这在为单细胞数据设计的方法中尤为明显。例如，对单细胞 GRN 推断方法的基准研究表明，这些方法在实验数据和硅学（模拟）数据上的准确性和共识性都很差，尤其是当推断过程中考虑的基因数量增加时更是如此。[24,108,109].不足为奇的是，与实验数据集相比，一些方法在应用于硅学数据集时表现更好，这可能是因为硅学网络与真正的生物 GRN 相比具有更简单的网络架构。[110].不过，由于缺乏黄金标准实验来确定基本真相，使用硅学 GRN 是一个很好的中间选择，目前也是验证和基准 GRN 推断方法的 流行策略。

硅学 GRN 作为地面实况模型的替代物的有效性取决于其准确模拟 TF、CRE 和基因之间复杂的直接和间接关系的能力。[23].这仍然是一个重大挑战，因为用于生成硅学 GRN 的基本假设往往是对真实生物网络中基本调控联系的过度简化。[110].除了 Li 及其同事最近提出的 一种多基因组 GRN 模拟方法（scMultiSim），旨在捕捉不同 omics 层（RNA 和 ATAC）之间的调控相互作用之外，目前还缺乏硅学多基因组 GRN。虽然这是朝着构建更具生物准确性的硅学 GRN 迈出的重要一步，但也存在一些重要的局限性，包括缺乏染色质可访问区域的输出。因此，基因与调控域之间没有联系，无法作为多基因组 GRN 推断方法的基准。此外，由于缺乏可访问的调控区域及其各自的序列，因此无法进行 TF 主题富集分析，以推断和验证重建 GRN 中的 TF-CRE 相互作用。

从另一个角度看，评估重建的 GRN 和基准 GRN 推断方法是密切相关的。一个可靠的模型应能有效捕捉到观测数据的 特征，并因此能够生成与基本事实非常接近的模拟数据。因此，就 GRN 推断而 言 ，一个有效的模型应能生成准确模拟 TF、CRE 和基因之间调控关系的数据。简而言之，能否生成稳健的硅学基因组网络取决于基因组网络推断方法能否忠实地模拟基本事实，而基本事实也可以由实验验证的知识来指导。目前无法做到这一点表明，GRN 推断的假设和方法还不足以捕捉 GRN 的真实复杂性。虽然所有模型都有其固有的局限性和假设，但我们建议研究人员考虑驱动其方法推断过程的 假设是否必要，是否具有生物学意义。这不仅能提高未来 GRN 推断方法的通用性和准确性，还能增强我们准确模拟单细胞多组学数据结构的能力。

结 论

单细胞多组学技术和 GRN 推断方法的 并行发展为全面描述细胞类型和细胞状态基因调控关系提供了独特的机会。随着可用数据复杂性的增加，人们开发出了更强大的 GRN 推断方法来利用这些数据。在这篇综述中，我们对最新的、最先进的 GRN 推

断方法进行了分类和总结。基于相关性的方法可捕捉线性（scMEGA、STREAM）或非线性（FigR、TRIPOD）成对调控因子，并对其进行分析。

关系。同样，基于回归的方法可利用线性模型（Pando、scREMOTE、RENIN）或非线性模型（DIRECT-NET、SCENIC +）确定能解释目标基因表达的关键 TFs。概率模型（scMTNI）可结合先验 信息，确定每个基因最可能的调控因子。基于动态系统的方法（Dictys）结合外部 nal 因素来模拟基因表达随时间的变化。最后， 深度学习方法使用人工神经网络发现不同 omics 层 （DeepMAPS、MTLRank、LINGER）之间复杂的调控关系。

基因组网络推断是一个充满活力、发展迅速的研究领域，最近涌现出的新的单细胞多组学基因组网络推断方法就是证明。技术进步和算法 微创新将继续推动更强大工具的开发，从而发现新的调控相互作用，这对理解驱动细胞特性和疾病的调控网络起着至关重要的作用。不过，虽然目前的 GRN 推断方法比以前的 方法更先进，但仍有许多工作要做，以减少目前的局限性，提高 推断 GRN 的稳健性和准确性。尽管如此，单细胞测序技术和GRN推断方法显然都取得了 ，并将继续发展，以进一步准确重建多模式调控关系，这将对包括健康和疾病在内的广泛研究领域产生 。

收到：接收：2023 年 8 月 14 日；接受：2023 年 10 月 2 日；
Published online: 19 October 2023

参考文献

1. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease.*Cell*.152, 1237-1251 (2013).
2. Lambert, S. A. et al.*Cell*.172, 650-665 (2018).
3. Spitz, F. & Furlong, E. E. M. 转录因子：从增强子结合到 发育控制。*Nat.Rev. Genet*.13, 613-626 (2012).
4. Almeida, N. et al. 利用核心调控回路确定细胞身份。*EMBO J*. (2021). https://onlinelibrary.wiley.com/doi/10.15252/embj.2020106785.
5. Karlebach, G. & Shamir, R. 基因调控网络的建模与分析。*Nat.Rev. Mol*.9, 770-780 (2008).
6. Bar-Joseph, Z. et al. 基因模块和调控 网络的计算发现。*Nat.Biotechnol*.21, 1337-1342 (2003).
7. Ruan, J., Dean, A. K. & Zhang, W. A general co-expression network-based approach to gene expression analysis: comparison and applications.*BMC Syst*.4, 8 (2010).
8. Song, L., Langfelder, P. & Horvath, S. 共同表达量的比较：互信息、相关性和基于模型的指数。*BMC Bioinform*.13, 328 (2012).
9. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.*Nat.Methods* 10, 1213-1218 (2013)。
10. Lieberman-Aiden, E. 等人. 长程相互作用的综合图谱 揭示了人类基因组的折叠原理。*科学》* 326 卷，289-293 页（2009 年）。
11. Robertson, G. 等人. 使用 染色质免疫沉淀和大规模平行测序分析 STAT1 DNA 关联的全基因组图谱。*Nat. ods* 4, 651-657 (2007).
12. Lee, J., Hyeon, D. Y. & Hwang, D. Single-cell multiomics: technologies and data analysis methods.*Exp.Exp.Exp*.52, 1428-1442 (2020).
13. Tang, F. et al. 单细胞的 mRNA-Seq 全转录组分析。*Nat.方法* 6, 377-382 (2009)。
14. Nagano, T. et al. 单细胞 Hi-C 揭示了染色体 结构的细胞间变异性。*自然》* 502, 59-64 (2013)。
15. Rotem, A. et al. 单细胞 ChIP-seq 揭示了由 染色质状态定义的细胞亚群。*Nat.Biotechnol*.33, 1165-1172 (2015).
16. Cha, J. & Lee, I. 用单细胞网络生物学解决细胞异质性问题人类疾病。*Exp.Exp.Exp*.52, 1798-1808 (2020).
17. Zhang, L., Zhang, J. & Nie, Q. DIRECT-NET：从单细胞多组学数据中发现顺式调控元件并构建调控网络的高效方法。8, eabl7393 (2022).

18. Zhang, S. Y. & Stumpf, M. P. H. Learning cell-specific networks from dynamical single cell data. 预印本 https://doi.org/10.1101/2023.01.08.523176 (2023).
19. Ogbeide, S., Giannese, F., Mincarelli, L. & Macaulay, I. C. Into the multiverse: advances in single-cell multiomic profiling.Trends Genet.*TIG* 38, 831-843 (2022).
20. Ma, S. et al. 通过共享单细胞 RNA 和染色质图谱鉴定染色质潜力。*细胞》* 183, 1103-1116.e20 (2020).
21. Stoeckius, M. 等人. 在单细胞中同时测量表位和转录组。*Nat.Methods* 14, 865-868 (2017).
22. González-Blas, C. B. et al. SCENIC +: single-cell multiomic inference of enhancers and gene regulatory networks.*Nat.Methods* 20, 1355-1367 (2023).
23. Tran, A., Yang, P., Yang, J. Y. H. & Ormerod, J. T. scREMOTE: Using multimodal single cell data to predict regulatory gene relationships and to build a computational cell reprogramming model.*NAR Genomics Bioinform.*4, lqac023 (2022).
24. Chen, S. & Mar, J. C. 评估推断基因调控网络的方法，凸显其在单细胞基因表达数据方面的不足。*BMC Bioinform.*19, 232 (2018).
25. Mercatelli, D., Scalambra, L., Triboli, L., Ray, F. & Giorgi, F. M. 基因调控网络推断资源：实用概述。*Biochim.Biochim.Acta BBA - Gene Regul.Mech.*1863, 194430 (2020).
26. Badia-i-Mompel, P. 等人. 单细胞多组学时代的基因调控网络推断。*Nat.Rev. Genet.*24, 739-754 (2023).
27. Zuin, J. 等人. 通过增强子-启动子相互作用的非线性转录控制。*自然》* 604 卷，571-577（2022 年）。
28. 单细胞时代的特征选择重温。 *Genome Biol.* 22, 321 (2021)。
29. Huynh-Thu, V. A. & Sanguinetti, G. Gene regulatory network inference: an introductory survey：*方法与协议》*（Sanguinetti, G. & Huynh-Thu, V. A. 编辑）。1-23 (Springer, 2019). https://doi.org/ 10.1007/978-1-4939-8882-2_1.
30. Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E. & Guthke, R. Gene regulatory network inference：动态模型中的数据整合--综述。*Biosystems* 96, 86-103 (2009).
31. Polynikis, A., Hogan, S. J. & Di Bernardo, M. 比较基因调控网络的不同 ODE 建模方法。*J. Theor.*261, 511-530 (2009).
32. Yaghoobi, H., Haghipour, S., Hamzeiy, H. & Asadi-Khiavi, M. A review of modeling techniques for genetic regulatory networks.*J. Med.*2, 61-70 (2012).
33. Min, S., Lee, B. & Yoon, S. 生物信息学中的深度学习。*Brief.Bioinform.*18, 851-869 (2016).
34. Sapoval, N. et al. 在生物科学领域应用深度学习的当前进展与挑战。*Nat.Nat.*13, 1728 (2022).
35. Cao, Y., Geddes, T. A., Yang, J. Y. H. & Yang, P. Ensemble deep learning in bioinformatics.*Nat.Mach.Intell.*2, 500-508 (2020).
36. Liu, C., Huang, H. & Yang, P. 利用 Matilda 从多模态单细胞 omics 进行多任务学习。*Nucleic Acids Res.* 51, e45 (2023).
37. 深度学习塑造单细胞数据分析。*Nat.Rev. Mol.*23, 303-304 (2022).
38. Song, Q., Ruffalo, M. & Bar-Joseph, Z. Using single cell atlas data to reconstruct regulatory networks.*Nucleic Acids Res.* 51, e38 (2023).
39. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. 利用互补 DNA 微阵列定量监测基因表达模式。*科学》* 270 卷，467-470 页（1995 年）。
40. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns.*Proc.Natl Acad.*95, 14863-14868 (1998).
41. Faith, J. J. et al. 从表达谱简编中大规模绘制和验证大肠杆菌转录调控。*PLoS Biol.* 5, e8 (2007).
42. Margolin, A. A. et al：哺乳动物细胞背景下基因调控网络的重建算法。*BMC Bioinforma.*7, S7 (2006).
43. Meyer, P. E., Kontos, K., Lafitte, F. & Bontempi, G. Information-theoretic inference of large transcriptional regulatory networks.*EURASIP J. Bioinforma.Syst.Biol.* 2007, 1-9 (2007).
44. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. 使用基于树的方法从表达数据推断调控网络。*PLoS ONE* 5, e12776 (2010).
45. Wagner, A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.*Bioinformatics* 15, 776-784 (1999).
46. Kamal, A. et al：增强子介导的基因调控网络的推断与评估。*Mol.Syst.*19, e11627 (2023).
47. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data.*Proc.Natl.*114, E4914-E4923 (2017).

48. Duren, Z., Chen, X., Xin, J., Wang, Y. & Wong, W. H. 基于配对表达和染色质可及性数据的时程调控 分析。*Genome Res.* 30, 622-634 (2020).
49. Lemmens, K. et al. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli.10, R27 (2009).
50. 转录因子的 ChIP-Seq 预测 胚胎干细胞中的绝对和差异基因表达。*Proc.Natl Acad.*106，21521-21526 (2009)。
51. Wang, P. et al. ChIP-Array 2: integrating multiple omics data to construct gene regulatory networks.*Nucleic Acids Res.* 43, W264-W269 (2015)。
52. Jiang, Y. 等. 转录因子、靶基因和顺式调控 区域之间的三 关系的非参数单细胞多组表征。*Cell Syst.*13, 737-751.e4 (2022).
53. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics.*Nature* 590, 300-307 (2021)。
54. Kim, H. J. 等人. 综合统计学习揭示多能性进展过程中的转录网络动态。*Nucleic Acids Res.* 48, 1828-1842 (2020)。
55. Meng, C., Kuster, B., Culhane, A. C. & Gholami, A. M. A multivariate approach to the integration o f multi-omics datasets.*BMC Bioinforma.*15, 162 (2014).
56. Jerby-Arnon, L. et al. 癌细胞程序促进T细胞排斥和 检查点阻断的 抗性。*Cell* 175, 984-997.e24 (2018)。
57. Segerstolpe, Å. 等人.健康和 2 型糖尿病人胰腺 胰岛的单细胞转录组图谱分析. 细胞代谢研究.2010 年.*Cell Metab.*24, 593-607 (2016).
58. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age.*Nat.Commun.*11, 4307 (2020).
59. Vento-Tormo, R. 等人 单细胞重建人类早期母胎 界面。*Nature* 563, 347-353 (2018)。
60. Vieira Braga, F. A. 等人.人类肺部细胞普查发现健康和哮喘中的新型细胞 状态 .*Nat.Med.*25, 1153-1163 (2019).
61. Osorio, D., Zhong, Y., Li, G., Huang, J. Z. & Cai, J. J. scTenifoldNet：从单细胞数据构建和比较全转录组基因 调控网络的机器学习工作流。*Patterns* 1, 100139 (2020).
62. Matsumoto, H. et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation.*Bioinformatics* 33, 2314-2321 (2017)。
63. Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures.*Cell Syst.*5, 251-267.e3 (2017).
64. Li, H. et al. 基于图神经网络从单细胞 ATAC- seq 数据推断转录因子调控网络。*Nat.Mach.Intell.*4, 389-400 (2022).
65. Jiang, J. et al. IReNA: Integrated regulatory network analysis of single-cell transcriptomes and chromatin accessibility profiles. *iScience* 25, 105359 (2022).
66. Kamimoto, K. 等人. 通过网络推断和硅学 基因扰动剖析细胞身份。*Nature* 614, 742-751 (2023)。
67. Alanis-Lobato, G. et al. MICA: A multi-omics method to predict gene regulatory networks in early human embryos. 预 印 本 ： https://doi.org/10.1101/2023.02.03.527081 (2023)。
68. Jansen, C. 等人. 利用关联自组织图从 scATAC-seq 和 scRNA- seq 构建基因调控网络。*PLOS Comput.*15, e1006555 (2019).
69. Duren, Z. 等人用 scREG 对单细胞多组基因表达和 染色质可及性数据进行调控分析。*Genome Biol.* 23, 114 (2022).
70. Jin，S.、Zhang，L. & Nie，Q. ScAI：一种用于综合分析并行单细胞转录组和表观基因组图谱的无监督方法。*基因组* Biol. 21, 25 (2020).
71. Zeng, W. et al. DC3 是一种从 批量和单细胞基因组学数据中进行解卷积和耦合聚类的方法。*Nat.Nat.*10, 4613 (2019).
72. Cao, Z. J. & Gao, G. 多组学单细胞数据整合及图联嵌入的调控 推断。*Nat.Biotechnol.*40, 1458-1466 (2022).
73. Lin, Y. et al. scTIE: data integration and inference of gene regulation using single-cell temporal multimodal data. 预 印 本 ： https://doi.org/10.1101/2023.05.18.541381 (2023)。
74. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. 单细胞和空间多组学的方法与应用。*Nat.Rev. Genet.*24, 494-515 (2023).
75. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell.*Nat.Biotechnol.*37, 1452-1457 (2019).
76. Zhu, C. et al. 小鼠大脑单个 细胞组蛋白修饰和转录组的联合分析。*Nat.Methods* 18, 283-292 (2021)。
77. Mimitou, E. P. et al. 单细胞中染色质可及性、 基因表达和蛋白质水平的可扩展多模式分析。*Nat.Biotechnol.*39, 1246-1258 (2021).
78. Yuan, Q. & Duren, Z. 通过对非配对观测数据的回归分析整合单细胞多组学数据 。*基因组生物学》*，23, 160 (2022).

79. Kartha, V. K. et al.
omics。*Cell Genomics* 2, 100166 (2022).
80. Wang, L. et al. Dictys：动态基因调控网络利用单细胞多组学剖析发育的连续性。*Nat.Methods* 20, 1368-1378 (2023).
81. Steinacher, A., Bates, D. G., Akman, O. E. & Soyer, O. S. 基因调控中的非线性动力学促进了基因表达水平的稳健性和可进化性。*PLOS ONE* 11, e0153295 (2016).
82. Fleck, J. S. et al.*自然》*，621，365-372（2023 年）。
83. Ledru, N. et al.通过单细胞多组测序的正则回归分析预测上皮细胞状态的调节因子。预印本：https:// doi.org/10.1101/2022.12.29.522232 (2022)。
84. Zhang, S. et al. 从单细胞奥米克数据集推断细胞系的细胞类型特异性基因调控网络。*Nat.14, 3064 (2023).*14, 3064 (2023).
85. De Torrenté, L. et al. 基因表达分布的形状很重要：结合分布形状如何改进癌症转录组学数据的解读。*BMC Bioinforma.*21, 562 (2020).
86. Van Dongen, S. 贝叶斯统计中的先验规范： Three cautionary tales.*J. Theor.*242, 90-100 (2006).
87. Ma, A. et al. 使用异构图转换器的单细胞生物网络推断。*Nat.Nat.*14, 964 (2023).
88. Yuan, Q. & Duren, Z. 利用图集级外部数据对单细胞多组数据的基因重组建模进行持续终身学习。预印本：https://doi.org/10.1101/2023.08.01.551575 (2023)。
89. La Manno, G. 等人 单细胞的 RNA 速度。*Nature* 560, 494-498 (2018)。
90. Hicks, S. C., Townes, F. W., Teng, M. & Irizarry, R. A. Missing data and technical variability in single-cell RNA-sequencing experiments.*Biostatistics* 19, 562-578 (2018).
91. Deaton, A. M. et al. 免疫系统中基因内 CpG 岛的细胞类型特异性 DNA 甲基化。*Genome Res.* 21, 1074-1086 (2011)。
92. Ding, J. et al. 单细胞和单核 RNA 测序方法的系统比较。*Nat.Biotechnol.*38, 737-746 (2020).
93. Jiang, R., Sun, T., Song, D. & Li, J. J. Statistics or biology: the zero-inflation controversy about scRNA-seq data.*Genome Biol.* 23, 31 (2022).
94. Van Dijk, D. 等人. 利用数据扩散从单细胞数据中恢复基因相互作用.*Cell* 174, 716-729.e27 (2018).
95. Loney, T. & Nagelkerke, N. J. The individualistic fallacy, ecological studies and instrumental variables: a causal interpretation.*Emerg.Themes Epidemiol.*11, 18 (2014).
96. Steel, D. G. & Holt, D. Analysing and Adjusting Aggregation Effects：再论生态谬误。Int. Stat.Stat.Rev. Rev. *Int. Stat.*64, 39 (1996).
97. Hou, W., Ji, Z., Ji, H. & Hicks, S. C. A systematic evaluation of single-cell RNA-sequencing imputation methods.*Genome Biol.* 21, 218 (2020).
98. Sekula, M., Gaskins, J. & Datta, S. A sparse Bayesian factor model for the construction of gene co-expression networks from single-cell RNA sequencing count data.*BMC Bioinforma.*21, 361 (2020).
99. Altman, N. & Krzywinski, M. Association, correlation and causation.*Nat.方法* 12, 899-900 (2015).
100. Pearl, J. Statistics and causal inference：A review.*Test* 12, 281-345 (2003).
101. Meinshausen, N. 等人. 从基因扰动实验和验证中推断因果关系的方法。*Proc.Natl Acad.*113, 7361-7368 (2016).
102. Qiu, X. et al. 使用 Scribe 从耦合单细胞表达动态推断因果基因调控网络。*Cell Syst.*10, 265-274.e11 (2020).
103. Streit, A. et al：小鸡作为模型系统：*Genesis* 51, 296-310 (2013)。
104. Tegnér, J., Yeung, M. K. S., Hasty, J. & Collins, J. J. Reverse engineering gene networks：基因扰动与动态建模的整合。*Proc.Natl Acad.*100, 5944-5949 (2003).
105. Akinci, E., Hamilton, M. C., Khowpinitchai, B. & Sherwood, R. I. Using CRISPR to understand and manipulate gene regulation.*Development* 148, dev182667 (2021).
106. Xie, S., Duan, J., Li, B., Zhou, P. & Hon, G. C. Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells.*Mol. Cell 66, 285-299.Cell* 66, 285-299.e5 (2017).
107. 徐伟等人的 ISSAAC-seq 使单细胞染色质可及性和基因表达的多模式分析变得灵敏而灵活。*Nat.方法* 19, 1243-1249 (2022)。
108. Kang, Y., Thieffry, D. & Cantini, L. Evaluating the reproducibility of single-cell gene regulatory network inference algorithms.*Front.Genet.*12, 617282 (2021).
109. Nguyen, H., Tran, D., Tran, B., Pehlivan, B. & Nguyen, T. A comprehensive survey of regulatory network inference methods using single cell RNA sequencing data.*Brief.Bioinform.*22, BBAA190 (2021).
110. Pratapa, A., Jalihal, A. P., Law, J. N., Bharadwaj, A. & Murali, T. M. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data.*Nat.Methods* 17, 147-154 (2020).

致谢

衷心感谢每位作者的以下资助来源以及手稿的编写工作：D.K.由澳大利亚联邦政府研究培训计划津贴奖学金和儿童医学研究所补足奖资助；A.T.由澳大利亚联邦政府研究培训计划津贴奖学金资助。J.Y.H.Y. 和 P.Y. 由香港创新科技署的 AIR@innoHK计划资助。P.Y.获得了澳大利亚国家健康与医学研究委员会（NHMRC）的研究者资助（1173469）和澳大利亚国家干细胞基金会（National Stem Cell Foundation of Australia）的梅特卡夫奖（Metcalf Prize）。作者感谢悉尼精准数据科学中心（Sydney Precision Data Science Centre）同事的支持和智慧讨论。他们还要感谢 Farhan Ameen、Tom Geddes、Carissa Chen、Lijia Yu、Jackson Zhou 和 Helen Fu 的反馈和编辑。

作者贡献

P.Y.和J.Y.构思了这项研究。D.K.和A.T.负责审 稿 ，H.K.和Y.L.参与撰写。

利益冲突

作者声明不存在利益冲突。

其他信息

通讯和资料索取请联系 Jean Yee Hwa Yang 或 Pengyi Yang。