自然 评论 遗传学
https://doi.org/10.1038/s41576-023-00618-5

评论文章

检查更新

单细胞多组学时代的基因调控网络推断

wangjiaji
2024-07-08 09:11:57

染色质、转录因子和基因

Pau Badia-i-Mompel[1], Lorna Wessels[1,2], Sophia Müller-Dott[1], Rémi Trimbour[1,3], Ricardo O. Ramirez Flores[1], Ricard Argelaguet[4] & Julio Saez-Rodriguez[1]

摘要

染色质、转录因子和基因之间的相互作用产生了复杂的调控回路，可以用基因调控网络（GRN）来表示。研究基因调控网络有助于了解细胞特性是如何建立、维持和在疾病中被破坏的。基因调控网络可以从实验数据（历史上的大容量全局数据）和/或文献中推断出来。

单细胞多组学技术的出现促进了新型计算方法的发展，这些方法利用基因组、转录组和染色质可及性信息，以前所未有的分辨率推断 GRN。在此，我们回顾了从转录组学和染色质可及性数据中推断转录因子与基因相互作用的 GRN 的关键原则。我们重点对使用单细胞多模态数据的方法进行比较和分类。我们强调了 GRN 推断所面临的挑战，特别是在基准测试方面，以及使用其他数据模式的进一步发展潜力。

章节

[1]海德堡大学医学院、海德堡大学医院、计算生物医学研究所、Bioquant，德国海德堡。[2]血管生物学和肿瘤血管生成系，欧洲血管科学中心，海德堡大学医学院，德国曼海姆。[3]法国巴黎，巴黎城市大学巴斯德研究所，CNRS UMR 3738，整合基因组学机器学习小组。[4]Altos Labs, Granta Park, Cambridge, UK.✉ 电子邮件：pub.saez@uni-heidelberg.de

导言

细胞通过调节基因转录来协调细胞活动，以响应细胞内外的信号。转录主要由转录因子（TFs）调控，这些蛋白质与 DNA 的特定序列（DNA 结合位点）结合，可对目标基因的转录率产生积极或消极的影响。[1].基因组 DNA 与结构蛋白紧密结合成称为核小体的复合体，核小体是染色质的基本单位，因此转录机制无法接触到大多数基因。为了实现转录，基因转录起始位点附近的区域（称为启动子）需要通过移位紧密排列的核小体来暴露出来。DNA 可及性的改变可由所谓的先驱 TFs 结合引发。[2].其他 TF 可与 DNA 的远端*顺式调控*元件（CRE）结合，并与辅助因子和其他蛋白质一起，共同促成 RNA 聚合酶蛋白复合物的招募和稳定，从而从基因体 DNA 合成 mRNA（图 1a）。

基因调控网络（GRNs）是以网络（数学上也定义为图形）为形式的基因表达调控的可解释计算模型。基因调控的多种成分，如TFs、剪接因子、长非编码RNAs、微RNAs和代谢物，都可以纳入基因调控网络。在此，我们重点讨论其最简单的表示方法，即只捕捉 TF 与目标基因之间的相互作用，GRN 的节点由基因组成，其中一些是 TF，GRN 的边代表基因之间的调控相互作用（图 1b）。其他可能的 GRN 表示法将在其他地方讨论[3–6].揭示 GRN 的拓扑结构和动力学是理解细胞特性如何建立和维持的基础，这对细胞工程的设计具有重要意义。[7]这对细胞命运工程[8]和疾病预防具有重要意义。[9].

了解 GRNs 是生物学领域的一项长期探索，20 世纪 60 年代描述细菌乳糖（lac）操作子的开创性工作就证明了这一点。[10].利用各种高通量实验方法和计算算法，重建大规模基因组网络成为系统生物学的一个主要焦点[11–13].从历史上看，GRN 通常是由数据库中经过实验验证的调控事件组装而成的[14–17]或从大量转录组学数据中的基因共表达中推断出来的[18–20].如果有足够的转录组学数据，就可以推断出比从数据库中提取的 GRN 更符合当前生物学问题背景的 GRN，后者往往具有普遍性。然而，转录组学数据并不能直接捕捉到许多潜在的调控机制，如 TF 蛋白丰度和 DNA 结合事件、TF 与辅助因子的合作、替代转录本剪接、翻译后蛋白质修饰事件以及基因组的可及性和结构。纳入并测量基因调控的这些其他方面有可能生成更能代表体内基因调控的 GRN（图 1b）。例如，加入染色质可及性数据[21]数据可通过考虑基因是否开放以及在推断 GRN 时纳入 CRE 来微调 TF 与基因的联系。

此外，批量分析提供的是组织样本中各细胞类型的混合测量结果，因此无法区分特定细胞类型或细胞状态的调控方案[22,23].单细胞技术的使用克服了这一局限性[24,25]单细胞技术的使用克服了这一局限性，可以推断不同细胞类型、分化轨迹和条件下的 GRN（图 1c）。因此，随着多模态图谱技术的引入[26–28]因此，随着多模态图谱技术的引入，最近出现了大量新型 GRN 推断方法。

wangjiaji
2024-07-08 09:19:54

转录组学数据

在这篇综述中，我们概述了 GRN 推断的一般原则及其潜在的局限性。此外，我们还介绍了如何利用多模态读出结果来推断更准确的 GRN，并对针对这一任务开发的几种新型工具进行了分类和简要介绍。此外，我们还强调了可能的下游 GRN 分析以及如何通过实验评估所获得的结果。最后，我们讨论了该领域当前面临的挑战和未来的发展方向。

全球资源网络的推断

全球基因网络推断是指利用计算方法，将基因调控这一高度复杂的动态过程归纳为可解释的数据网络结构的过程。它基于这样一个假设，即可以从分子数据中观察和测量真正的潜在 GRN 的效应（图 1b）。[29](图 1b）。遗传资源网络中的相互作用可以是有向或无向的（分别表示基因之间存在因果关系或不存在因果关系）、有符号的（表示调控模式，正向或负向）和/或加权的（表示相互作用的强度）。

来自转录组学数据

这一类方法是根据其他基因的表达情况来建立模型，试图解释观察到的基因表达变异性。加权基因共表达网络分析（WGCNA）[19]是最简单、最流行的方法之一。它在整个转录组中进行成对相关性分析，以确定共表达基因的模块。由于相关性的对称性，由此产生的网络通常被称为基因共表达网络，其相互作用是不定向的。虽然这种策略有助于以无监督的方式识别基因模块，但由于缺乏因果调控联系，其可解释性受到影响，通常会产生大量假阳性关联。为了解决这些局限性，GENIE3（参考文献 20）及其更快的实现方法 GRN-Boost2（参考文献 30）等方法首先根据先前报告的调控活动将 TF 与目标基因区分开来，然后训练模型来预测基因的表达。[31]然后仅根据 TFs 的表达来训练预测目的基因表达的模型，这就大大减少了需要考虑的相互作用的数量。通过这种方法，无向相互作用变成了有向连接，从而引入了推测的因果关系。然而，仅从转录组学数据进行推断会产生假阳性，因为许多参与基因调控的其他机制，如染色质可及性，都被忽略了。此外，由于编码 TF 的 mRNA 转录本要成为功能性蛋白质需要许多过程，因此仅靠转录本水平可能信息量不够大。[1,32].这些局限性可能会阻碍推断过程，因为有研究表明，总体而言，这些方法在准确推断 GRN 方面的成功率一般[33–35].

来自 TF 结合数据或染色质可及性

染色质免疫共沉淀测序（ChIP-seq）、靶标下裂解和标记（CUT&Tag）等检测方法[36]以及靶标下裂解和标记（CUT&Tag）[37]等检测方法可以测量整个基因组的 TF 结合情况。通过将 TF 结合位点分配给推测的靶基因，这些信息可直接用于构建 GRN。[38].然而，尽管有一些高通量替代方法[39–41]但是，尽管有一些高通量的替代方法，TF 结合的分析仍然成本高昂，而且仅限于有抗体可用的 TF。此外，仅使用 TF 结合数据通常需要将结合的 TF 按最接近的基因组位置分配到其靶基因上，而忽略了已知与基因调控相关的可能的远端相互作用事件。[1].相比之下，一项开创性的研究探索了将 ChIP-seq

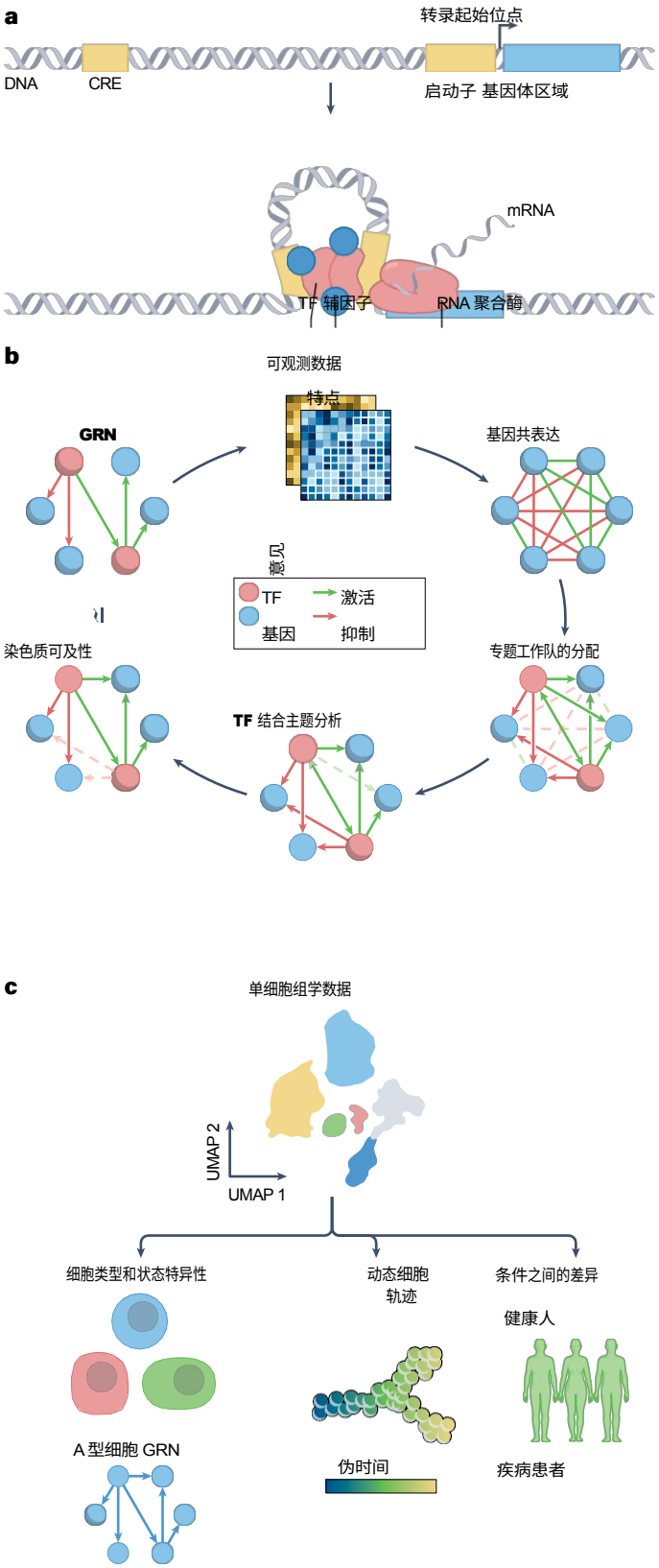


图 1 基因调控网络的原理 a, 基因调控及其关键要素。转录因子（TFs）与启动子区域和*顺式调控*元件（CREs）结合，置换核小体，使启动子区域和*顺式调控*元件（CREs）与核小体结合。b. 基因调控网络TFs、辅助因子和其他蛋白质之间的合作使 RNA 聚合酶蛋白复合物得以招募和稳定，从而从基因体 DNA 合成 mRNA。**b**, 基因调控网络（GRNs）可以从测量的 omics 数据中推断出来，通过对 TF 结合预测或染色质可及性等附加信息的建模，可以完善基因调控网络，使其更接近真正的基础 GRN。GRN 的节点是 TF 及其调控基因，节点之间的边表示调控模式（激活或抑制）。c. 通过单细胞全息数据生成的 GRN 可以了解细胞类型和状态的特异性，解释动态轨迹的进展，并识别不同条件下的差异。

数据和转录组学数据，从而能够更精细地将 TFs 分配到不依赖于最接近基因的基因上。[42].

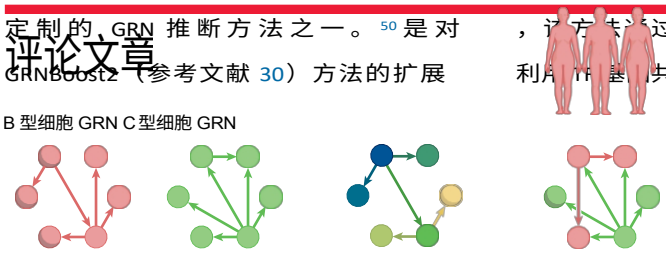
另一种方法是利用染色质可及性数据来推断可能成为 TF 靶标的基因调控元件。最常用的技术是转座酶染色质可及性测序法（ATAC-seq），因为其操作简单，成本相对较低。[21]但也有其他技术，如 DNase-seq[43]和 NOME-seq[44](等技术（综述见[45]).利用染色质可及性数据的方法将 GRN 推断分为两步：第一步，将 TFs 分配到基因调控元件（开放染色质区域，通常称为峰）；第二步，将这些调控元件分配到基因（图 2）。第一步，利用 TF 结合基调数据库和基调匹配算法，对可访问的 CRE 上的 TF 进行结合预测（方框 1）。第二步，方法将可访问的 CRE 链接到一定基因组距离内的基因。距离截止值是基于这样的观察：远端 CRE（如增强子或沉默子）通常与基因启动子区域的相互作用距离通常为[1].此类推断方法的一些实例包括 ATAC2GRN（参考文献 46）、LISA[47]和 SPIDER[48].这些方法假定，如果基因的启动子区域可被访问，则该基因正在被转录，但事实可能并非总是如此。

来自单细胞转录组学数据

利用大容量组学数据进行 GRN 推断的方法能够描述全基因组调控事件的特征，但在组织等混合样本的情况下，这些方法无法捕捉 GRN 的细胞类型或状态特异性。[22,23].此外，GRN 推断方法需要大样本量才能产生足够的数据，而这在大样本分析中成本过高。

随着单细胞技术，特别是单细胞 RNA 测序（scRNA-seq）技术的出现，GRN 重构方法已被用于推断细胞类型特异性 TF 基因相互作用，以及这些 GRN 在不同发育阶段和条件下发生的动态变化（图 1c）。[49](图 1c）。SCENIC 是最早为 scRNA-seq 数据

定制的 GRN 推断方法之一。[50] 是对 GRNBoost2（参考文献 30）方法的扩展，该方法通过利用 TF 基因共表达模式生成细胞类型特异性 GRN，此外还根据 TF 结合基团富集情况修剪 GRN 边缘



基因启动子区域。单细胞测量分辨率的提高还有助于识别细胞的动态状态及其转变，这些状态可能不容易区分为不同的组别，例如在发育、细胞分化或疾病进展过程中[51,52].伪时间排序表征了这些连续的

wangjiaji
2024-07-08 10:06:36

顺式调控元件

wangjiaji
2024-07-08 09:33:09

单细胞转录组学数据

变化，并可用于为全球资源网络推断提供信息。由此得出的全球资源网络为了解关键命运决策所涉及的复杂过程提供了宝贵的信息。LEAP[53]和 SINCERITIES[54]是利用伪时间排序推断 GRN 中基因之间方向性的 GRN 推断方法的实例。使用差异测试后获得的对比度统计[55,56]是识别不同情况（如健康个体与疾病患者群）之间差异的有效方法。这种策略不同于计算 GRN 之间的差异，这在后面描述 "下游 GRN 分析 "的章节中会有解释。

单细胞染色质可及性分析（如单细胞ATAC-seq（scATAC-seq））的最新进展[57]可与单细胞转录组学一起进行[26–28]可与单细胞转录组学一起进行，从而以无与伦比的清晰度完善了基因组网络的重建。早期的一些研究从未配对的多组学数据中推断出 GRNs，用于研究人类髓系细胞分化、小鼠胚胎发育和发育过程。[58]小鼠胚胎发育[59]和树突状细胞的 HIV 感染[60].然而，他们并没有将自己的方法作为工具提供给他人使用。随后，利用 scRNA-seq 和 scATAC-seq 进行 GRN 推断的新方法层出不穷（表 1 和图 2；见图 3）。

补充方框 1）。用于推断 GRN 的多模态数据如果来自同一细胞，则可以配对；如果来自不同细胞，则可以不配对。有些方法不要求每个细胞的染色质可及性和基因表达谱相匹配，因为它们要么总结各组细胞的读出数据，要么为每种模式独立构建 GRN，然后再进行合并。相比之下，其他方法则在同一细胞中同时建立两种模式的模型。在这些 "同时 "方法中，如果使用整合方法对两种模式进行匹配，则仍可对未配对的数据进行建模。[61].为方便使用，其中一些方法（如 DeepMAPS[62]、FigR[63]、GLUE[64]、scAI[65]和 SOMatic[66]）采用了自己的集成方法。

多模态 GRN 推断方法使用单模态方法所使用的扩展框架来重建 GRN。具体而言，这些方法从 TF 基因表达预测基因表达，利用结合基序信息将 TF 分配到可访问的 CRE，并根据基因组距离将 CRE 与目标基因联系起来（图 2）。在预测 TF 结合事件时，不同的方法使用不同的、高度异质的 TF 结合基调数据库和预测算法（表 1 和方框 1）。由于 TF 结合基调数据库

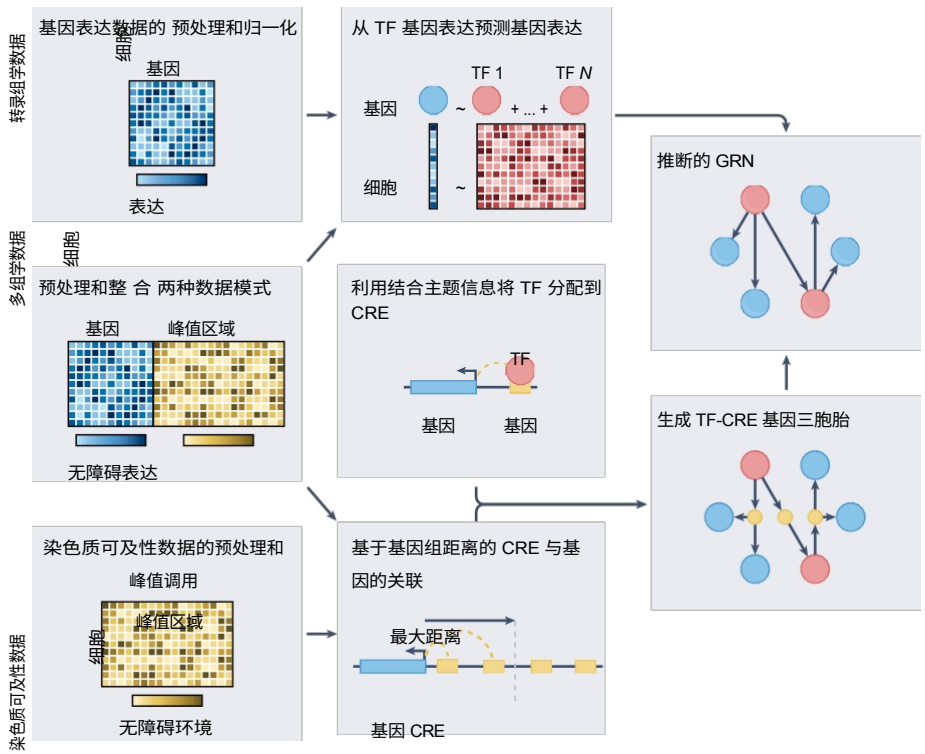


图 2：基因调控网络推断方法流程图。基因调控网络（GRN）推断方法涉及不同的步骤，具体取决于研究样本或细胞生成的数据模式。首先要对转录组学数据进行预处理和归一化处理，以建立基因表达矩阵，其中包含各基因在不同时间段的转录水平。

样本或细胞。从其他来源获取已知转录因子（TF）基因列表，以区分具有调控能力的基因。然后，通过建立模型来推断 TF 与目标基因之间的相互作用，这些模型试图通过 TF 转录本丰度来预测观察到的基因表达，从而产生 TF 基因关联。最后，将获得的相互作用汇总并表示为 GRN。首先对染色质可及性数据进行预处理

和峰值被调用，以建立一个峰值可达性矩阵，其中包含二进制的

有关不同样本或细胞中*顺式*调控元件（CRE）开放性的信息。根据基因组距离限制将 CRE 与基因联系起来，并利用 TF 结合基调数据库和基调匹配算法预测 TF 与 CRE 结合。这些信息被用于获得 TF-CRE 基因三联体。最后，这些相互作用被简化为 TF 基因对，并汇总到 GRN 中。当样本通过转录组学和染色质可及性（多组学数据）分析时，预处理如果需要，还可以整合未配对的模式。在两种模式都可用的情况下，各种方法可以同时利用上述三个建模步骤来构建 TF-CRE 基因三联体，然后将其简化并汇总到 GRN 中。

由于 TF 的覆盖范围不同，预测算法对结合的建模方式也不同，即使使用类似的建模策略，GRN 推断方法之间的结果也可能不同。大多数方法允许使用不同于其默认值的 TF 结合主题数据库，但大多数方法固定使用的主题匹配算法--除了 SCENIC+[67]除外，它实现了三种算法，即 cisTarget[67]、DEM[67]和 HOMER[68].此外，GRN 推断方法使用不同的基因组距离截断值将开放染色质区域分配给目标基因。有些方法考虑的近距离可达 10 kb，有些方法考虑的中等距离可 达 100 kb，有些方法考虑的大距离效应可达 1,000 kb，还有一些方法在原始出版物或源代码中都没 有 指定距离截止值（表 1）。鉴于功能验证的相互作用在最近的距离上大量富集，而在 100 kb 的距离上大幅下降[1,69]，距离截止值的不同很可能会影响推断出的 GRN。

在完成上述步骤（图 2）后，多模态 GRN 推断方法会生成一个候选支架网络，该网络由与目标基因相连的 CRE 相关的 TF 三部分组成。要生成最终的 GRN 结构，需要使用不同的数学策略。其中一些策略假定 TF、CRE 和基因之间是线性关系，另一些则假定是非线性关系（表 1）。线性建模假设一个变量（如基因转录本）的变化与另一个变量（如 TF 转录本或 CRE 开放度）成正比。相比之下，非线性建模可以考虑变量之间更复杂的相互作用，如协同效应[70].尽管人们普遍认为基因表达是一个非线性过程[70]尽管人们普遍认为基因表达是一个非线性过程，但由于基因表达网络的线性建模在表述和解释上较为简单，因此通常更受青睐。无论使用哪种建模策略，都可以使用频繁法或贝叶斯概率统计框架来评估所获得的调控相互作用的意义（表 1）。频数法将事件的概率定义为该事件在大量相同实验中发生的次数比例，而贝叶斯概率则将其定义为根据观测数据和以往信息对上述事件发生的置信度。贝叶斯方法可以考虑现有的先验知识，但与频数法相比，贝叶斯方法通常需要更多的计算资源，这在利用大规模单细胞数据推断全基因组 GRN 时可能会受到限制。此外，贝叶斯推断的成功与否还取决于所使用的先验知识的质量。因此，当没有先验信息或怀疑先验信息不准确时，频数推断可能会更准确。

多模态 GRN 推断方法可根据其建模策略和接受的输入类型进行分组（表 1）。大多数方法旨在通过频数回归对不同组别（通常是细胞类型）的 GRN 进行建模。FigR[63]和 GRaNIE[71]等使用频谱线性回归；DIRECT-NET[72]和 SCENIC+[67]使用频数主义非线性回归（随机森林）；而 PECA[73]和 Symphony[74,75]使用贝叶斯模型。相比之下，CellOracle[76]、Inferelator 3.0（参考文献 77）和 Pando[78]则为用户提供了多种建模策略。如果由于数据的连续性（例如在细胞发育过程中）而无法从数据中定义不同的组别，scMEGA[79]和 IReNA[80]利用轨迹分别以线性和非线性方式推断 GRN。此外，Dictys[81]、scMTNI[82]和 TimeReg[83]结合使用细胞类型分组和轨迹数据为 GRN 模型提供信息，而 CellOracle[76]和 SCENIC+[67]则使用后者进行下游分析。ANANSE[84]、sc-compReg[85]和 SCENIC+[67]构建

结合主题数据库和主题匹配算法

生成多个转录因子（TF）的全基因组结合数据需要费力的实验，因此基因调控网络（GRN）推断方法转而根据先验信息预测开放基因组区域的 TF 结合事件。这些信息来自大量的 TF-DNA 结合实验，如染色质免疫共沉淀后测序（ChIP-seq）实验。[36]可用于提取特定 TF 最可能特异性结合的基因组序列，即通常所说的 TF 结合基序[208]。有几个数据库收集了此类检测结果，并生成了模式生物的 TF 结合基调集。由于不同数据库的覆盖范围可能不同，因此可以合并这些数据库，以增加 GRN 推断过程中考虑的 TF 数量。此外，还开发了几种利用 TF 结合基团的计算算法来预测结合事件，即所谓的主题匹配算法。所有这些算法都是根据主题序列计算 TF 结合事件的概率，并筛选出重要的结合事件。由于不同的方法对 TF 结合的建模方式不同，因此它们之间的结果可能会有差异，在进行 GRN 推断时应加以考虑。下表列出了TF结合主题数据库和主题匹配算法。

名称	网址	参考文献
结合图案数据库		
CIS-BP	http://cisbp.ccbr.utoronto.ca/	209
cisTarget 数据库	https://resources.aertslab.org/cistarget/数据库/	67
ENCODE	https://www.encodeproject.org/software/encode-motifs/	210
HOCOMOCO	https://hocomoco11.autosome.org/	211
JASPAR	https://jaspar.genereg.net/	212
TRANSFAC	https://genexplain.com/transfac/	213
UniPROBE	http://thebrain.bwh.harvard.edu/uniprobe/	214
动机匹配器算法		
FIMO	https://snystrom.github.io/memes-manual/	215
GimmeMotifs	https://gimmemotifs.readthedocs.io/	216
荷马	http://homer.ucsd.edu/homer/motif/	68
情绪（如在 motifmatchr 中实现）	https://github.com/jhkorhonen/MOODS https://github.com/GreenleafLab/motifmatchr	217, 218
motifanalysis （在 reg-hint 中实现）	https://reg-gen.readthedocs.io/	219
PIQ 工具包	https://bitbucket.org/thashim/piq-single/src/master/	220
PWMScan	https://ccg.epfl.ch/pwmtools/pwmscan.php	221
pycisTarget	https://pycistarget.readthedocs.io/	67

表 **1** 从多组学数据推断基因调控网络的现有工具

工具[a]	可能的输入	多模式数据类型	建模类型	互动类型	统计框架	默认图案数据库/图案匹配器	默认的上游/下游距离截断点	语言	参考文献
ANANSE	组别、对比	不配对	线性	加权	常客	CIS-BP/ GimmeMotifs	100 kb/100 kb	Python	84
细胞甲骨文	群体、轨迹	不配对	线性	签名，加权	频繁论或贝叶斯论	CIS-BP/ GimmeMotifs	500 kb/500 kb	Python	76
DC3	组别	不配对	线性	二进制	常客	未定义/ HOMER	基于 Hi-C	Python	88
深度地图	组别	配对或集成	线性	加权	常客	JASPAR/ PWMScan	150 kb/150 kb 或外显子	Python	62
Dictys	群体、轨迹	非配对/配对或综合	线性	签名，加权	常客	霍莫科/霍默	500 kb/500 kb	Python	81
直接网络	组别	配对或集成	非线性	二进制	常客	JASPAR/MOODs	250 kb/250 kb	R	72
FigR	组别	配对或集成	线性	签名，加权	常客	CIS-BP/MOODs	50 kb/50 kb	R	63
胶水	组别	配对或集成	非线性	加权	常客	JASPAR/cisTarget	150 kb/150 kb	Python	64
GRaNIE	组别	配对或集成	线性	加权	常客	贾斯帕，霍霍莫科/ PWMscan	250 kb/250 kb	R	71
干扰器 2.5	组别	非配对	线性	签名 加权	频数 或贝叶斯	CIS-BP, ENCODE, TRANSFAC/FIMO	10 kb/10 kb	蟒蛇	203
干扰器 3.0	组别	非	配对线性或非线性	加权	频数法 或贝叶斯	JASPAR/FIMO	50 kb/2.5 kb	蟒蛇	77
IReNA	轨迹	不配对	线性	签名，加权	常客	TRANSFAC/FIMO	250 kb/250 kb	R	80
神奇	组别、对比	不配对	非线性	加权	贝叶斯	CIS-BP, ENCODE/ MOODs	基于 Hi-C	R/MATLAB	166
多国部队	组别	不配对	非线性	签名，加权	常客	贾斯帕尔-霍霍莫科动机分析	最近的誊写起始点	R	204
潘多	组别	配对或集成	线性或非线性	签名，加权	频繁论或贝叶斯论	JASPAR, CIS-BP/ MOODs	100 kb/基因体	R	78
PECA	组别	配对或集成	线性	加权	贝叶斯	JASPAR, TRANSPAC、UniPROBE/ HOMER	1 000 kb/1 000 kb	MATLAB	73
监管主题	组别	配对或集成	线性	已签署	常客	HOCOMOCO/ 动机分析	5 kb/5 kb	MATLAB	205
雷宁	组别	配对或集成	线性	签名，加权	常客	CIS-BP/MOODs	500 kb/500 kb	R	206
scAI	组别	配对或集成	线性	加权	常客	CIS-BP/MOODs	250 kb/250 kb	R	65
sc-compReg	组别、对比	不配对	线性	二进制	常客	未定义/未定义	未定义	R	85

SCENIC+	群体、对比、轨迹	配对或集成	线性	签名，加权	常客	cisTarget/ cisTarget	150 kb/150 kb	Python	67
scMEGA	轨迹	配对或集成	线性	加权	常客	JASPAR/MOODs	250 kb/250 kb	R	79
scMTNI	群体、轨迹	不配对	线性或非线性	加权	贝叶斯	CIS-BP/PIQ	5 kb/5 kb	C++	82
SOMatic	组别	不配对	线性	二进制	常客	HOCOMOCO FIMO	50 kb/50 kb	C++	66

表 1（续） | 从多组学数据推断基因调控网络的现有工具

工具[a]	可能的输入	多模式数据类型	建模类型	互动类型	统计框架	默认图案数据库/图案匹配器	默认的上游/下游距离截断点	语言	参考文献
交响乐	组别	不配对	线性	签名，加权	贝叶斯	未定义/FIMO	最近的誊写起始点	Python	74,75
时间管理	群体、轨迹	配对或集成	线性	二进制	常客	未定义/ HOMER	未定义	MATLAB	83
三脚架	组别	配对或集成	非线性	签名，加权	频繁论或贝叶斯论	HOCOMOCO, JASPAR/MOODs	100 kb/100 kb	R	207

[a]有关这些工具及其方法的更多详情，见补充方框 1。

在推断过程中，不仅可以利用基因组特异性（如细胞类型特异性）GRN，还可以利用基因对比统计。

GRN 下游分析

一旦从任何分辨率和 omics 数据组合中推断出 GRN，就可以利用各种下游分析对其进行查询，从而提供新的生物学见解（图 3 和方框 2）。

拓扑分析

尽管基因调控网络是一种简单且可解释的基因调控模型，但它仍可能包含大量基因以及更多基因之间的相互作用。网络中心性度量有助于确定哪些 TF 或基因对网络的连通性或信息流更为重要（图 3a）。网络中心性度量的一些例子包括度中心性、接近中心性、间度中心性和特征向量中心性。These measures have been useful to identify TFs that drive cell fate changes in diverse biological contexts, such as direct lineage reprograming[76]人类心肌梗塞[86]和小鼠发育[87].

另一种表征GRN拓扑结构的方法是使用基于谱图理论的方法，这种方法可以探索以矩阵形式表示的网络的适当联系。例如，将非负矩阵因式分解应用于GRN的邻接矩阵，发现了在小鼠胚胎干细胞中协同驱动系转的TF组[83,88].同样，GRN 拓扑聚类确定了人类造血细胞分化中的已知调节因子[70]和巨噬细胞对干扰素-γ 的反应中的已知调节因子。[71].然后，可以对获得的基因调控模块进行基因集富集，以确定其潜在生物功能的特征[89].

比较分析

对 GRNs 进行比较分析可以发现驱动细胞类型、细胞状态、疾病状态、治疗方法和生物体之间差异的重配事件（图 3b）。最简单的比较分析方法是成对地减去 GRN 之间 TF 基因的相互作用。这种方法确定了淋巴细胞白血病患者 B 细胞亚群中的关键调控因子[85]发现了将成纤维细胞转分化为不同人类细胞类型的 TFs 群组[84]发现了阿尔茨海默病特异性候选跨*调节因子*[90]和人类 T 细胞中细胞状态特异性调节因子[74,75].它还被用于评估TF-基因相互作用的进化保守性和转录调控在不同物种间的适应性。[91].然而，由于 GRNs 的稀疏性和嘈杂性，直接比较 TF 基因的相互作用往往不够稳健。

主题建模策略，如潜在德里赫特分配（latent Dirichlet allocation），是一种无监督贝叶斯模型，最初是为自然语言处理开发的。[92]这种方法可以生成高密度、低维度的表征，过滤 GRN 结构中的噪声，从而更稳健地捕捉调控关系中的差异。这种策略有助于预测癌症患者的生存期[93]以及识别人类造血过程中的重新布线事件。[82].

推断专题工作队的活动

GRNs 可与富集方法相结合，从转录组学数据中推断 TF 的激活关系[15,50,94,95].这种方法可以将观察到的基因表达与 GRN 拓扑整合起来，从而提取出哪些 TF 在某些情况下可能具有相关作用（图 3c）。常见的富集方法包括 GSEA[96]、AUCell[50]和 VIPER[94]等。[95].在批量研究中，通过富集方法推断 TF 活性，可以识别可用药的肿瘤蛋白、对药物治疗有反应的细胞系分层等。[94]细胞系对药物治疗的反应分层[97]以及确定一个主调节因子作为乳腺癌的转移促进因子。[98].在单细胞研究中，富集方法确定了人类 T 细胞[99]、少突胶质细胞瘤的调节因子和诱导因子以及治疗少突胶质细胞瘤的潜在药物靶点。[50]以及 COVID-19 患者病理成纤维细胞的潜在药物靶点（参考文献 100）。这些方法最近还被应用到空间分辨转录组学数据中，例如提出了参与人类心肌梗塞缺血性病变周围边界区心肌细胞功能转换的调控因子。[86].

细胞命运的干扰和预测

通过以迭代方式将 TF 表达传播到目标基因，GRN 可用于模拟随时间变化的基因表达值。有了这个框架，就可以通过改变候选 TF 的表达来进行硅学扰动，然后观察迭代一定次数后对转录组的影响（图 3d）。之后，可将模拟值与局部相邻细胞的基因表达量进行比较，以估算细胞身份转换概率，类似于 RNA 速度分析[101]。这种策略首先由 CellOracle[76]首次提出，这一策略表明了 *Zfp57* 在生成和维持小鼠诱导内胚层祖细胞中的作用，后来体外扰动实验也验证了这一点。SCENIC+[67]使用类似的策略确定了 *RUNX3* 是黑色素细胞转化为间质黑色素瘤细胞的潜在驱动力，展示了 GRNs 捕捉和模拟复杂调控事件的能力。

全球资源网络的实验评估

GRN 推断方法预测的连接应被视为假设的调控相互作用，必须通过补充信息和/或实验进行评估。在本节中，我们将讨论这方面的常见做法（图 4）。

TF 丰度和翻译后修饰

编码特定 TF 的转录本数量只能有限地代表其蛋白质丰度，更不用说其活性了[102]。为此，蛋白质组学技术可用于测量 TF 的丰度。单细胞分辨率的焦油蛋白质组学仍具有挑战性，但有些

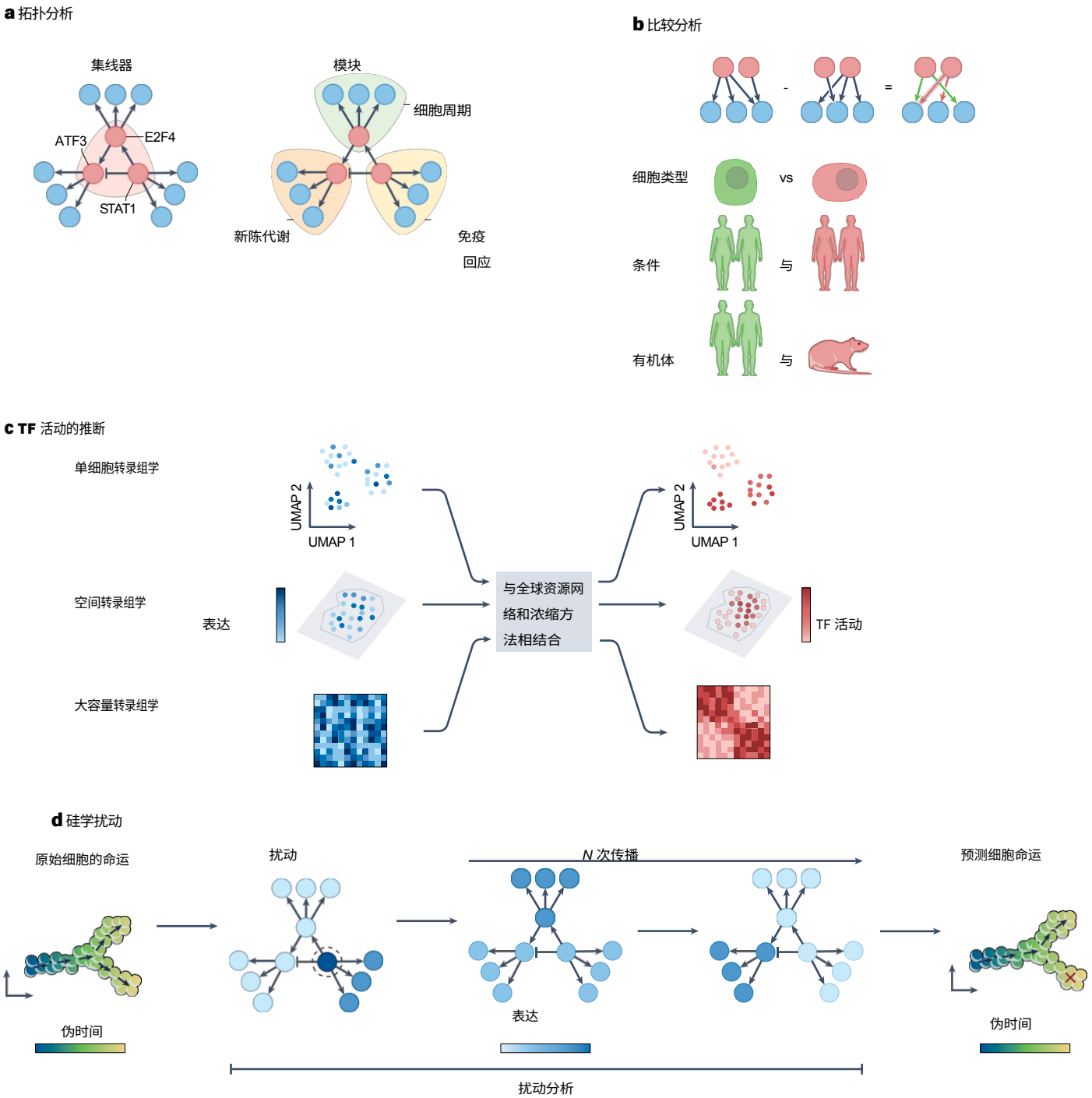


图 3 基因调控网络的应用 a, 拓扑分析。网络中心度量可用于识别基因调控网络（GRN）中高度连接的转录因子（TFs）或基因中心。根据节点的连通性对节点进行聚类，可以产生与生物功能相关联的子网络模块。通过成对减去 GRN 之间 TF 基因的相互作用，比较不同 GRN 的连通性，可以深入了解不同细胞类型、个体、条件之间基因调控的重新布线情况。

c，推断 TF 活性。GRNs 可与富集方法相结合，从转录组学数据中推断哪些 TF 可能具有功能活性。从多组学数据中推断出的 GRNs 可用于推断 TF 在其他情况下的活性，如独立的单细胞、空间或大容量转录组学数据。通过短时间的迭代将基因表达的变化传播到网络中，GRN 可用于模拟扰动实验。获得的模拟基因表达谱可用于推断细胞命运的决定。

方框 2

成功应用单细胞多组学数据衍生的基因调控网络

基因调控网络（GRNs）已被广泛用于回答一系列研究问题；在此，我们总结了最近的一些实例。

- 利用单细胞 RNA 测序（scRNA-seq）和单细胞转座酶染色质检测法，生成了人脑器官组织的多模态时间过程。
  测序（scATAC-seq）[78].作者利用GRN推断工具Pando预测了转录因子（TF）结合位点，并推断出了支撑类器官发育的全球GRN。他们通过基于CRISPR的筛选，对GRN进行了硅学预测，发现GLI3是皮层命运建立的重要转录因子。
- 通过以下方法生成了苍蝇大脑的多模态图谱
  作者利用 scRNA-seq 和 scATAC-seq 数据分析了神经元的发育、重编程和成熟轨迹[222]。作者利用在 omics 数据上训练的深度学习模型，推断出细胞类型特异的 TF 结合预测，并以此解码神经元多样性所依赖的增强子架构的调控语法。
- CellOracle 是一个数学模型，用于利用单细胞多组学数据训练 GRN，对 TF 进行硅学扰动。[76].在斑马鱼发育的背景下，作者对 TF 基因敲除进行了系统的预测，其中包括
  通过该研究，发现了斑马鱼早期发育的关键调控因子（包括 *noto* 和 *lhx1a*）*的*新作用。
- 通过分析单个细胞的基因表达和染色质可及性，建立小鼠早期器官形成的多模式图谱[87].作者开发了染色质免疫共沉淀-测序（ChIP-seq）技术，这是一种
  方法来预测 TF 结合位点，并用它来描述神经*母*细胞向体细胞中胚层过渡所依赖的 GRN。使用 CellOracle[76]框架进行硅学预测，然后进行实验验证，他们确定了 Brachyury 在启动顺式 *调控*元件（CRE）分化中的作用。
- 对阿尔茨海默病患者的皮层组织进行了 scRNA-seq 和 scATAC-seq 研究[90].通过模拟
  通过研究 TFs、CREs 和目标基因之间的关系，作者发现 ZEB1 和 MAFB 是参与基因调控的候选基因，它们有可能参与神经元和小胶质细胞的疾病进展。
- 作者使用 Inferelator 3.0（参考文献 77）推断了来自小鼠的几个 $CD4^+$ 记忆 T 细胞群的 GRNs，并通过收集 42 个已识别 TF 的 TF 基因敲除和 ChIP-seq 数据对结果进行了基准测试[223]。作者将获得的
  我们还从功能上验证了T滤泡辅助细胞中涉及IL-6、MAF和CD153的调控回路，这对老龄小鼠抗体介导的疫苗反应非常重要。

质谱法等技术或使用抗体-寡核苷酸共轭物的检测方法已可用于[103]（综述见[104]）。此外，还可查询人类蛋白质图谱[105,106]等数据库，以确认候选 TF 蛋白水平是否已在特定组织或细胞类型中出现过。此外，磷酸化、泛素化和甲基化等翻译后修饰可影响 TF 的定位、稳定性、活性以及与其他蛋白质的相互作用[107]。研究得最多的 TF 翻译后修饰是磷酸化，磷酸化可以说明 TF 是处于非活性还是活性状态。[108]

TF 结合与合作

GRN推断方法依赖于基于结合主题分析的TF结合预测，将TF分配到基因组中的开放染色质区域。众所周知，这种类型的预测会产生许多假阳性，因为大量的 TF 结合主题具有低特异性[109]。为此，ChIP-seq[36]可用于测试 GRN 推断方法正确预测了多少 TF 结合事件。[76].ChIP- Atlas[110]、EpiMap[111]和 UniBind[112]等数据库汇集了大量 ChIP-seq 实验数据，并按生物体、组织和细胞类型进行分类，是分析 GRN 预测的宝贵资源。因为并非所有 ChIP-seq 峰都代表 TF 的直接结合事件、

BindingDB[111] 和 UniBind[112] 采用不同的策略来整理数据，从而提供有关 TF 结合位点的更可靠信息。另一种替代方法是单分子足迹法，这是一种以单 DNA 分子分辨率联合测量 TF 结合和核小体占位的技术[113,114] 。它可以检查每个基因组区域的状态频率：与 TF 结合、未结合但有开放的 chro- matin 或未结合但被核小体占据。与 ChIP-seq 相比，单分子足迹分析的优势在于它提供了 TF 结合的动态量化状态，而不是二元描述。GRN推断方法预测多个TF结合到同一开放基因组区域，这符合TF与DNA合作结合诱导转录的知识。[1,115,116].评估 GRN 的另一种方法是检查该网络是否重新发现了 TF 的合作结合。CAP-SELEX[117] 等技术可在 DNA 存在的情况下联合分析选定 TF 对之间的合作性相互作用。单分子足迹法也可用于此目的，检查足迹的重叠情况。其他方法包括使用蛋白质-蛋白质相互作用测定[118] 或检查以前注释过的相互作用数据库[119] 。

CRE 的监管活动

CRE 可处于三种不同的染色质状态：转录活跃、静止或抑制。许多已报道的开放染色质区域可能在基因调控中没有作用，从而增加了CRE的数量。

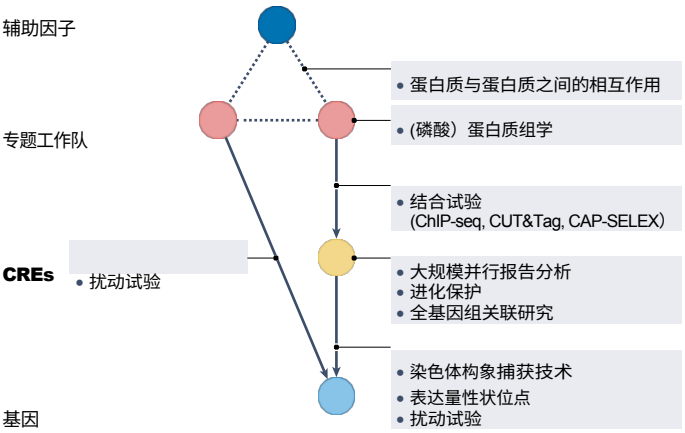


图 4 基因调控网络的实验评估。虽然基因调控没有明确的 "基本事实"，但可以通过一些实验和分析来验证基因调控网络（GRN）的特定方面。转录因子（TFs）之间的相互作用可在蛋白质-蛋白质数据库中查询到，TFs 有大量共同的靶基因，并假定它们之间存在相互作用。TF 蛋白的存在可以
蛋白质组测定可确认TFs的活化状态，靶向磷酸化实验可评估TFs的活化状态。TF与*顺式*调控元件（CRE）之间的联系可通过结合实验来确认，例如
染色质免疫沉淀后测序（ChIP-seq）、目标下裂解和标记（CUT&Tag）以及 CAP-SELEX。候选的 CREs 可通过报告实验或扰动实验来检测其调控能力。另外，还可以假定功能性 CREs 在进化过程中是保守的，或者富含在通过全基因组关联研究确定的疾病相关位点中。可以评估 CRE 与基因之间的联系
通过基因组构象测定（如 Hi-C 或超分辨显微镜）或基于 CRISPR 的扰动实验来实现。此外，还可以使用表达定量性状位点数据库。

因此，实验必须评估候选增强子是否会影响基因表达。因此，实验必须评估候选增强子是否会影响基因表达。大规模并行报告测定[120]是一种可以测试候选基因组区域是否能诱导外显子载体中基因表达的技术。另一种策略是对候选的 CREs 进行基于 CRISPR 的汇集分析，然后进行 RNA 测序，以确定基因组中哪些区域会影响基因表达。[69,121].此外，ENCODE 联合会利用各种生化检测方法，对 100 多万个候选 CREs 进行了编目，这些候选 CREs 具有类似增强子的特征，横跨人类基因组约 16% 的区域。[36,122,123].由于功能增强子在进化过程中是保守的[124,125]，另一种方法是检查不同物种的基因组序列相似性。在研究疾病时，可以对开放染色质区域的 CRE 进行扫描，以发现之前在全基因组关联研究中与疾病相关的单核苷酸多态性 (SNP)。如果候选 CREs 富集了与疾病相关的 SNPs，则表明这些 CREs 很可能具有功能性。[90,126].

与目标基因的联系

即使开放染色质区域被证实具有调控特性，也有必要检测它影响哪些特定基因。染色体构象捕获技术（如 Hi-C[127–129]）可以测量基因组区域之间的接触概率，并确定拓扑关联域。因为基因调控

基于持续的基因组相互作用[130,131]，如果一个基因组区域与一个基因的启动子区域持续保持密切接触，则该区域可能会调控该基因。为此，超分辨率显微镜也被用于验证候选的 CRE 基因相互作用[132]，尽管其通量低于 Hi-C。基于 CRISPR 筛选的扰动测定与全转录组分析相结合，可以删除或激活特定的 CRE，并观察其如何改变基因表达。[69,133].此外，人类群体中的表达定量性状位点数据库也可用于验证远端候选 CREs134-[136]。

TF 干扰

测试 TF 是否调控特定基因的一个更直接的方法是扰乱 TF 的表达，并观察其对基因表达的影响。目前已经有了集CRISPR（内）激活筛选与全转录组读出相结合的技术[137–141]。TF扰动后基因表达的变化可作为基本事实，以检查有多少受影响的基因被确定为 GRN 中的靶基因。[35,95,142].此外，我们还可以看到估计的 TF 活性（见上一节 "TF 活性推断"）是否与所进行的扰动（TF 的低或高活性分别表示 TF 的敲除或过表达）相对应。

挑战和未来方向

近年来omics数据集的积累，尤其是单细胞多omics数据的积累，推动了新一轮改进GRN推断策略的浪潮。加深对基因组网络的理解，不仅能利用这些模型理解基因调控的原理，还能将其作为驱动细胞工程中细胞命运决策的工具，从而生成具有新功能的新细胞类 型 ， 并将病变细胞重编程为健康的表型。前景非常广阔，但在 GRNs 建模和将其用作预测工具方面仍存在许多挑战。

转录与无障碍整合

多组学的使用原则上可以更好地反映基因调控，但也有其自身的挑战。作为高度关联的过程，染色质的可及性和转录在时间上是协调的。然而，它们的动力学有很大不同，并可能发生时间上的转移。人们通常认为，同一细胞在单个时间点的成对染色质可及性和转录组学数据代表了这两个过程的相互作用[143,144]。如果整合不充分导致 scATAC-seq 和 scRNA-seq 数据不匹配，会误导 GRN 的下游建模，那么这种局限性在未配对数据的情况下就会变得更加复杂[145]。新的整合策略，如 FigR 引入的策略，有望获得更好的匹配。[63]等新的整合策略有望在细胞间获得更好的匹配。除其他因素外，染色质可及性与基因表达之间的时间变化以及合作效应也会产生非线性关系。我们讨论过的一些 GRN 推理方法使用非线性公式来解释这一点，但与线性模型相比，它们失去了可解释性，而且往往不能明确捕捉到相互作用的符号。因此，出于计算可扩展性的考虑，许多方法仍然倾向于对基因调控进行线性建模。为了提高可解释性，SCENIC+[67]和 IReNA[80]首先使用随机森林非线性地推断调控相互作用，然后根据 TF 和基因转录本之间的相关性分析确定相互作用的符号。

术语表

通过测序分析转座可访问染色质（ATAC-seq）。一种用于识别利用超活性 Tn5 转座酶的转座可访问染色质区域的方法。

平均距离（最短路径的长度）的网络中心度量。计算一个节点的最短路径上的任何其他网络中的两个节点。

染色质的高阶丝状结构DNA 蛋白复合物，可存在于凝结或未凝结状态。

染色质免疫沉淀然后进行测序（ChIP-seq）。一种分析蛋白质与可访问的使用染色质免疫共沉淀技术检测 DNA 区域，然后进行 DNA 测序。

顺式*调节*元件（CRE）。非编码 DNA 区域调节附近基因与转录因子结合时因子 (TF)。这些因子包括启动子、增强器和消音器。

目标下的裂解和标记（CUT&Tag）。基于抗体的蛋白质分析技术利用转座酶与可访问的 DNA 区域相互作用Tn5 介导的标记，然后进行 DNA 测序。

接近中心度网络中心度量 描述节点到所有节点的最短路径的长度）。其他节点。

度数中心网络中心度量描述边的数量(度）。

DNA 结合位点转录的 DNA 序列因子可与基因结合，驱动基因监管，通常表现为核苷酸模式被称为图案。

特征向量中心性描述节点重要性的网络中心度量网络的邻国。

增强器远端 DNA 调控区域转录调控蛋白可以结合并激活转录。

表达量性状地点序列变异与基因表达变化相关的基因组位置。

基因调控网络（GRN）。基因调控网络（GRN转录调节因子和目标基因

全基因组关联研究分析方法，以确定常出现的单核苷酸基因组中的多态性横跨一个大群人。

Hi-C染色质研究技术在三维空间中的构象，以确定可能相距较远但在三维空间中更接近。

元胞分子结构相似的细胞群可汇总为一个单一的全息图，以减少数据的稀疏性。

动机匹配器算法字符串匹配算法检测DNA 中的转录因子结合点序列。

网络中心性一组图论度量的相对重要性。网络中的节点。

山峰构成表观遗传测序技术读数的可访问染色质区域。

促销员基因组中位于基因转录起始位点之前的调控区域。

消音器远端 DNA 调控区域转录调控蛋白可以结合并抑制转录。

单核苷酸多态性（SNP）。DNA 序列变异造成的。特定位置的核苷酸。

拓扑关联领域高交互频率的自交互基因组区域域内的 序列和与邻国相对隔离区域，形成三维染色体结构

转录因子（TF）。可改变通过与特定 DNA 结合实现转录序列。

单细胞数据的规模和稀疏性

GRN 推断方法需要大量的观测数据，以捕捉所研究生物过程的变异性。这些观察结果可以是单个细胞、样本或条件。单细胞技术可为给定样本生成成千上万的图谱，因此与批量图谱技术相比，更容易推断出更多生物背景下的遗传资源网络。然而，来自同一样本的细胞并不一定是独立的，不能被视为真正的生物复制品[146]。因此，要获得有意义的 GRNs，可能需要纳入不同的样本。此外，目前的单细胞基因组网络推断方法建立的是整个细胞群

的集合网络，并没有考虑到细胞可能来自不同样本。解决这一问题的一种候选方法是 LIONESS[147]，该方法在推断 GRN 时对每个样本的贡献进行建模，并能生成特定样本的调控相互作用。此外，单细胞数据本质上是稀疏和有噪声的，特别是

对于通过 scATAC-seq 获得的数据，需要使用适当的过滤器来确保最低质量[148,149]。对于成对的多组学技术，目前还没有一个系统的基准来比较它们与单组学对应技术的不同覆盖率和灵敏度[150]。虽然稀疏性是单细胞技术的一个已知特性[151,152]，但本文讨论的 GRN 推断方法都没有在建模时明确考虑稀疏性。有些方法采用数据转换来抵消这一限制。估算方法可用于减少 "遗漏"（由 mRNA 或可访问 DNA 读数的采样不足引起）的数量[153–155]，但已证明它们可能会对 GRN 重建产生不利影响[156]。据报道，将类似细胞聚合成伪大块图谱或元细胞[146,157,158] 的策略是有益的。[87].由于其稀疏性，大多数计算管道将 scATAC-seq 数据视为二进制数据，为每个细胞分配可访问或封闭的基因组区域。然而，众所周知，DNA 可及性的真实情况更为精细，可

涉及以动态方式波动的中间可及性区域[159]。因此，将染色质可及性数据视为二进制数据可能不利于下游分析[160,161]，而定量处理可及性的方法可能会改善 GRN 重建[154,155,162,163]。

三维基因组结构的调控作用

目前的 GRN 推断方法使用基于基因组距离的任意截断值将 CRE 分配给基因。这种过滤方法的目的是减少每个基因的搜索空间，从而减少所需的计算资源，并减少假阳性相互作用的数量。[69].不过，也有一些 CRE 与基因之间相互作用距离较远的例子，如 *MYC* 基因的增强子位于其下游近 2 Mb 处[164]。根据所使用的距离截断点，GRN 推断方法可能会错过关键的 CRE 基因相互作用。此外，有些相互作用是跨染色体发生的，如嗅觉受体选择过程中的报道[165]，而目前的 GRN 方法无法考虑这些相互作用。解决这一问题的方法之一是使用基于三维接近性的技术，如 Hi-C[127,129]，来评估 CRE 是否可能调控基因。DC3（参考文献 88）和 MAGICAL[166] 已成功应用了这一策略。尽管有一些高通量的替代方法[167,168]，但染色体构象捕获技术因其稀疏性[169]，以及仍需与其他模式整合和其方案难以复制等事实，带来了新的挑战[170,171]。在这些技术得到更广泛应用之前，人们一直使用计算方法来预测基于可及性数据（如 scATAC-seq 数据）的基因组三维结构[172,173]。在 GRN 建模中使用这些方法有可能克服使用基于距离的截止值的局限性。

改进 TF 结合预测

目前，GRN 推断方法将 TF 分配到 CRE 的策略依赖于 TF 结合主题数据库（方框 1）。每个数据库的主题集合覆盖范围不同，这可能会使预测结果产生偏差。基调数据库以 ChIP-seq 等前结合实验的数据为基础。然而，据估计，在人类基因组中编码的大约 1,600 个序列特异性 TFs 中，有 10% 没有可用的结合数据。[31,109].没有已知结合基调的 TF 被排除在 GRN 建模之外，而这一因素在非模式生物中更为严重，因为它们的已知 TF 结合基调往往少于其他研究更深入的生物。纳入缺失的 TF 的一个可能解决方案是在 GRN 推断过程中利用已知的蛋白质-蛋白质相互作用。此外，目前的 TF 结合基调是基于来自多种组织和细胞类型的数据。众所周知，TF 结合是一个高度特定的过程。[1]尽管现有的基因图谱仍与许多组织相关，但细胞类型特异性基因图谱模型可能有助于提高 TF 结合预测的准确性。最近基于深度学习的计算策略可以进行细胞类型特异性 TF 结合预测[174,175]。这些模型经过训练，可完全根据 DNA 序列预测细胞类型特异性 DNA 可及性。训练完成后，它们会通过硅突变或使用解释性机器学习策略（如 SHAP[176]），确定哪些核苷酸对可及性的影响最大。为了得出细胞类型特异的 TF 结合预测，这些方法将预测的核苷酸数量与结合图案相结合。虽然这些策略有可能更好地将 GRN 推断与背景联系起来，但它们需要使用大量数据对模型进行预训练，而且仍然局限于已知的 TF。

结合图案。随着联盟倡议的高质量细胞图谱的积累，我们设想这些策略最终能取代传统的 TF 结合基调预测。[49,177,178]我们预计，这些策略最终将取代传统的 TF 结合基调预测。此外，目前的 TF 结合预测是二元的，但定量定义可以提供更多信息。BANC-seq[179] 是一种定量测量 TF 结合亲和力的技术，有可能生成更准确的 GRN。

用于 GRN 推断的新兴多组学

转录组学和染色质可及性数据的配对分析可以更准确地推断基因组网络，但这种分析方法仍然成本高昂，限制了其广泛应用。较新的替代方法，如 ISSAAC-seq[180]，能以比商业 10× Multiome 试剂盒低得多的成本进行多组学分析。尽管如此，单独的 scRNA-seq 和 scATAC-seq 联合数据可能无法提供足够的信息来全面描述基因调控的特征。在这种情况下，包含更多数据模式的单细胞多组学分析技术的进步将至关重要[181]。NEAT-seq[182] 就是这类前景看好的技术之一，它能同时分析核内蛋白、染色质可及性和基因表达，通过纳入 TF 蛋白丰度，剔除 GRN 建模中可能出现的假阳性。另一个例子是 scChaRM-seq[183]，它同时分析 DNA 甲基化、染色质可及性和基因表达。它们的联合分析可根据甲基化状态对 TF 分配到 CRE 的位置进行微调。此外，ATAC-STARR-seq[184] 还能同时进行大规模并行报告分析和染色质可及性分析，以测试开放 CRE 的转录能力。非靶向单细胞蛋白质组学和磷酸化蛋白质组学的进步可能会使功能活跃的TFs图谱分析成为可能[185]。其中一个例子是磷酸质谱（Phospho-seq[186]），这是一种在单细胞水平分析染色质可及性和磷酸化蛋白的新技术。众所周知，个体群体间的遗传信息是异质的，但大多数方法都假定它们共享相同的基因组[187]。scGET-seq[188] 是一种联合记录基因组和染色质可及性的技术，通过测试 SNPs 如何因 TF 结合亲和力的变化而影响染色质可及性，有可能帮助推断因果关系 GRN。

全球资源网络的基准

要了解新型 GRN 推断方法的准确性，尤其是那些利用多组学数据的方法的准确性，GRN 的基准测试是至关重要的。遗憾的是，由于没有明确的基因调控 "基本真相"，因此对预测的基因遗传网络进行验证是一项复杂的任务。基准测试的一种方法是构建硅学 GRN，使我们能够根据已知的基本真相评估 GRN 重建。[34]但这可能并不能很好地反映真实的生物 GRN。如上一节所述，有不同的方法可用于间接评估预测基因调控事件的质量，但这些方法都有一定的局限性。即使观察到 TF 与某个基因结合，也不一定意味着 TF 对该基因进行了调控，因为 TF 与 DNA 开放区域的结合是随机的，需要与其他分子合作才能有效调控转录[159]。染色体构象捕获技术可提供接触信息并定义拓扑关联域。然而，其分辨率可能不足以检测某些基因组相互作用[189]。目前已有高分辨率的 Hi-C 图谱，如 Micro-C[190]，但在比较许多实验条件时，其成本会变得过高。为了解决这个问题，目前正在采用机器学习方法

用于从较低覆盖率数据推算较高覆盖率的 Hi-C 地图，以提高其分辨率[191]。另一种可能性是使用基于超分辨率显微镜的替代方法，但其通量相当有限[189]。TF 协同驱动基因表达，但它们主要是通过 DNA 介导的相互作用而非蛋白质-蛋白质接触实现的[117]。因此，对 TF-TF 相互作用的评估可能仅限于特定情况。由于其固有的因果关系，通过扰动实验评估 GRN 是一种更有前景的方法。然而，扰动筛选成本高昂，有时无法达到预期效果，而且可能会受到补偿机制和未计下游效应的影响。除了所有这些局限性之外，由于基因调控是一个时间依赖过程，实验可能会因为捕捉到不同的时间框架或实验噪音而自相矛盾。由于基因调控的真正 "金标准 "目前还无法产生，我们更倾向于将这些不同的评估策略作为 "银标准 "的集合。我们认为，收集和发布此类信息的计算工具将有助于社区对推断出的基因调控网络进行质量控制，并对新的基因调控网络推断方法进行基准测试。单细胞分析开放问题项目（Open Problems for Single-Cell Analysis）等平台[192]为运行和评估各种 GRN 推断方法提供了合适的基础设施。这些平台还能通过公开竞争，以无偏见的方式对 GRN 推断方法进行评估，DREAM 挑战项目对批量转录组学数据中的 GRN 推断方法就是一个很好的例子。[33].

全球资源网络

必须牢记，基因组网络并不是孤立的。漆操作子（lac operon）这一经典例子表明，代谢物（乳糖）会触发基因调控，这突出表明 GRNs 是纠缠在一起的细胞机制的一部分，其中包括信号传递和代谢过程。单细胞磷酸蛋白组学和代谢组学的加入[193]，为利用特定情境网络模型将基因调控与细胞信号过程联系起来提供了可能[194]。

此外，细胞很少作为独立系统工作，基因调控在组织内部高度协调。因此，多模态数据与空间信息的整合将是另一个大有可为的方向。特别是，我们设想将 GRN 与细胞内和细胞间通信过程[195–197]整合为空间感知模型[198,199]。这些策略有助于理解多细胞的时空调控过程。[200]

结论

高通量单细胞多模态技术和计算方法的进步正在为建立越来越精确的基因表达网络推断模型铺平道路。数据集的大规模使得从测序数据中训练深度学习方法来预测基因表达越来越成为可能[175,201,202]。GRN 通过提供更易解释的模型，对这些方法进行了补充。这些不同的方法合在一起，可以帮助我们更好地理解不同细胞类型、器官、种群和物种之间基因调控的差异，并作为控制细胞命运决定的工具。在生物医学领域，这些知识可以帮助确定控制不同疾病病理生理过程的新型药物靶点。

在线出版：2023 年 6 月 26 日

1. Lin, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code. *Mol. Cell* **83**, 373-392 (2023).*Cell* **83**, 373-392 (2023).
这篇内容广泛的综述涵盖了顺式*调节*代码的分子基础。
2. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: establishing competence for gene expression.*Genes Dev.***25**, 2227-2241 (2011).
3. Lai, X., Wolkenhauer, O. & Vera, J. 通过数学建模理解 microRNA 介导的基因调控网络。*Nucleic Acids Res.* **44**, 6019-6035 (2016).
4. Du，J.-X. et al. 剪接因子：对其在癌症替代剪接中的调控网络的见解。*Cancer Lett.***501**, 83-104 (2021).
5. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding RNAs and its biological functions.*Nat.Rev. Mol. Cell Biol.***22**, 96-118 (2021).
6. Carthew, R. W. Gene regulation and cellular metabolism: an essential partnership. *Trends Genet.***37**, 389-400 (2021).
7. Davidson, E. H. & Erwin, D. H. Gene regulatory networks and the evolution of animal body plans.*Science* **311**, 796-800 (2006).
8. Su, E. Y., Spangler, A., Bian, Q., Kasamoto, J. Y. & Cahan, P. Reconstruction of dynamic regulatory networks reveals signaling-induced topology changes associated with germ layer specification.*Stem Cell Rep.* **17**, 427-442 (2022).
9. Claringbould, A. & Zaugg, J. B. Enhancers in disease: molecular basis and emerging treatment strategies.*Trends Mol.Med.***27**, 1060-1073 (2021).
10. Jacob, F. & Monod, J. 蛋白质合成的遗传调控机制。*J.Mol. Biol.***3**, 318-356 (1961).
这项开创性的研究描述了一个基因调控系统。
11. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu.Rev. Genomics Hum.Genet.***2**, 343-372 (2001).
12. Davidson, E. H. et al.*Science* **295**, 1669-1678 (2002).
13. Snyder, M. & Gallagher, J. E. G. Systems biology from a yeast omics perspective. *FEBS Lett.***583**, 3895-3899 (2009).
14. Han, H. et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions.*Nucleic Acids Res.* **46**, D380-D386 (2018).
15. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities.*Genome Res.* **29**, 1363-1375 (2019).
16. Liu，Z.-P.，Wu，C.，Miao，H. & Wu，H. RegNetwork：人和小鼠转录和转录后调控网络的集成数据库。*Database* **2015**, bav095 (2015).
17. Keenan, A. B. et al. ChEA3：通过正交组学 整合进行转录因子富集分析。*Nucleic Acids Res.* **47**, W212-W224 (2019).
18. Margolin, A. A. et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.*BMC Bioinforma.***7**, S7 (2006).
19. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis.*BMC Bioinforma.***9**, 559 (2008).
20. Huynh-Thu, V. A., Irrthum, A., Wehenkel, L. & Geurts, P. Inferring regulatory networks from expression data using tree-based methods.*PLoS ONE* **5**, e12776 (2010).
21. 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.方法* **10**, 1213-1218 (2013)。
22. Fiers, M. W. E. J. et al.
Brief.Funct.Genomics **17**, 246-254 (2018)。
23. Cha, J. & Lee, I. Single-cell network biology for resolving cellular heterogeneity in human diseases.*Exp.Exp.Exp.***52**, 1798-1808 (2020).
24. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells.*Cell* **161**, 1187-1201 (2015)。
25. Macosko, E. Z. et al. 使用纳升液滴对单个细胞 进行高并行全基因组表达谱分析。*Cell* **161**, 1202-1214 (2015)。
26. Chen, S., Lake, B. B. & Zhang, K. 同一细胞中转录组和染色质可及性的高通量测序。*Nat.Biotechnol.***37**, 1452-1457 (2019).
27. Liu, L. 等人 单细胞多组学层的解卷积揭示了调控 的异质性。*Nat.Nat.***10**, 470 (2019).
28. 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).
29. Mercatelli, D., Scalambra, L., Triboli, L., Ray, F. & Giorgi, F. M. Gene regulatory network inference resources: a practical overview.*Biochim.Biochim.Acta Gene Regul.Mech.***1863**, 194430 (2020).
30. Moerman, T. et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks.*Bioinformatics* **35**, 2159-2161 (2019).
31. Lambert, S. A. et al.*Cell* **175**, 598-599 (2018).
32. Holland, C. H. et al. 转录因子和通路 分析工具在单细胞 RNA-seq 数据上的稳健性和适用性。*Genome Biol.* **21**, 36 (2020).
33. Marbach, D. et al. Wisdom of crowds for robust gene network inference.*Nat.方法* **9**、796-804 (2012).
这项工作是从批量转录组学 数据中推断 **GRN** 的众包基准。
34. 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).

35. McCalla, S. G. 等. 从单细胞 RNA-seq 数据确定计算 网络推断方法的优缺点。*G3* **3**, jkad004 (2023).
36. Johnson，D. S.，Mortazavi，A.，Myers，R. M. & Wold，B. 蛋白质-DNA 相互作用的全基因组图谱。*Science* **316**, 1497-1502 (2007).
37. Kaya-Okur, H. S. et al. CUT&Tag for efficient epigenomic profiling of small sample and single cells.*Nat.Commun.***10**, 1930 (2019).
38. Lee, T. I. et al. *酿酒酵母*的转录调控网络。*科学* **298**, 799-804 (2002).
39. Grosselin, K. 等人. 高通量单细胞 ChIP-seq 鉴定乳腺癌染色质状态的异质性。*Nat.Genet.***51**, 1060-1066 (2019).
40. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues.*Nat.Biotechnol.***39**, 825-835 (2021)。
41. Bartosovic, M. & Castelo-Branco, G. 使用基于纳米体的单细胞 CUT&Tag 进行多模式染色质谱分析。*Nat.https://doi.org/10.1038/s41587-022-01535-4* (2022).
42. Qin, J., Hu, Y., Xu, F., Yalamanchili, H. K. & Wang, J. Inferring gene regulatory networks by integrating ChIP-seq/chip and transcriptome data via LASSO-type regularization methods.*Methods* **67**, 294-303 (2014).
43. Boyle, A. P. et al. 全基因组开放染色质 的高分辨率图谱和特征。*Cell* **132**, 311-322 (2008).
44. Kelly, T. K. et al. 单个 DNA 分子内核糖体定位和 DNA 甲基化的全基因组图谱 .*Genome Res.* **22**, 2497-2506 (2012)。
45. Minnoye, L. et al. 染色质可及性分析方法。*Nat.Rev. Methods Prim.***1**, 1-24 (2021).
46. Pranzatelli, T. J. F., Michael, D. G. & Chiorini, J. A. ATAC2GRN: optimized ATAC-seq and DNase1-seq pipelines for rapid and accurate genome regulatory network inference.*BMC Genom.***19**, 563 (2018).
47. Qin, Q. et al. Lisa: 通过公共染色质可及性和 ChIP-seq 数据的整合建模 推断转录调节因子。*Genome Biol.* **21**, 32 (2020).
48. Sonawane, A. R., DeMeo, D. L., Quackenbush, J. & Glass, K. 利用表观遗传学数据构建基因调控 网络。*NPJ Syst.Biol.***7**, 45 (2021).
49. Tabula Sapiens Consortium et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans.*Science* **376**, eabl4896 (2022).
50. Aibar, S. et al. SCENIC：单细胞调控网络推断与聚类。*Nat.Methods* **14**, 1083-1086 (2017)。
这项研究首次提出了在单细胞水平上推断 **GRN** 的定制方法，，引入了利用 **TF** 结合基序信息估算 **GRN** 的方法。
51. Herring, C. A., Chen, B., McKinley, E. T. & Lau, K. S. Single-cell computational strategies for lineage reconstruction in tissue systems.*Cell Mol.Gastroenterol.Hepatol.***5**, 539-548 (2018)。
52. Wagner, A., Regev, A. & Yosef, N. 利用单细胞 基因组学揭示细胞身份的载体。*Nat.Biotechnol.***34**, 1145-1160 (2016).
53. Specht, A. T. & Li, J. LEAP: 使用伪时间排序为单细胞 RNA-sequencing 数据构建基因共表达网络。*Bioinformatics* **33**, 764-766 (2017).
54. Papili Gao, N., Ud-Dean, S. M. M., Gandrillon, O. & Gunawan, R. SINCERITIES: inferring gene regulatory networks from time-stamped single cell transcriptional expression profiles.*Bioinformatics* **34**, 258-266 (2018).
55. Love, M. I., Huber, W. & Anders, S. 使用 DESeq2 对 RNA-seq 数据的折叠变化和离散度 进行调节估算。*Genome Biol.* **15**, 550 (2014).
56. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies.*Nucleic Acids Res.* **43**, e47 (2015)。
57. Buenrostro, J. D. et al. 单细胞染色质可及性揭示了调控变异的原理。*Nature* **523**, 486-490 (2015)。
本文介绍了单细胞转座染色质检测（**scATAC**） 技术。
58. Ramirez, R. N. et al. 人类髓系分化的动态基因调控网络。*Cell Syst.***4**, 416-429.e3 (2017).
59. Starks, R. R., Biswas, A., Jain, A. & Tuteja, G. 对不同启动子可及性和基因表达谱的联合分析确定了组织特异性基因和主动 抑制网络。*Epigenetics Chromatin* **12**, 16 (2019).
60. Johnson, J. S. et al. 单核细胞衍生的树突状细胞先天感应 HIV 时转录网络的综合图谱。*Cell Rep.* **30**, 914-931.e9 (2020)。
61. Argelaguet, R., Cuomo, A. S. E., Stegle, O. & Marioni, J. C. 单细胞数据整合中的计算原理与 挑战。*Nat.Biotechnol.***39**, 1202-1215 (2021).
62. Ma, A. et al. 使用异构图 变换器进行单细胞生物网络推断。*Nat.Nat.***14**, 964 (2023).
63. Kartha, V. K. et al. 利用单细胞多组学进行基因调控的功能推断。*Cell Genom.***2**, 100166 (2022).
本文介绍了 **FigR**，它对 **scRNA-seq** 和 **scATAC-seq** 数据有一种新颖的整合策略，可以增强 **GRN** 推断能力。
64. Cao, Z.-J. & Gao, G. 多组学单细胞数据整合与调控推断（ graph-linked embedding）。*Nat.Biotechnol.***40**, 1458-1466 (2022).
65. Jin, S., Zhang, L. & Nie, Q. ScAI: an unsupervised approach for the integrative analysis of parallel single-cell transcriptomic and epigenomic profiles.*Genome Biol.* **21**, 25 (2020)。
66. Jansen, C. et al. 利用链接自组织图从 scATAC-seq 和 scRNA-seq 构建基因调控网络。*PLoS Comput.***15**, e1006555 (2019).

67. González-Blas, C. B. et al. SCENIC+：增强子和基因调控网络的单细胞多组学推断。Preprint at *bioRxiv* https://doi.org/10.1101/2022.08.19.504505 (2022).
这项研究展示了大量经过整理的 **TF** 结合基元，并介绍了一种新颖的 **GRN** 推理方法。
68. Heinz, S. et al. 巨噬细胞和 B 细胞特性所需的线型决定转录因子的简单组合为*顺式*调节元件。*Mol. Cell 38, 576-589 (2010).Cell* **38**, 576-589 (2010)。
69. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens.*Cell* **176**, 1516 (2019).
70. Zuin, J. et al. 通过增强子-启动子相互作用对转录的非线性控制。*自然》* **604** 卷，571-577 页（2022 年）。
71. Kamal, A. et al. GRaNIE and GRaNPA: inference and evaluation of enhancer-mediated gene regulatory networks.*Mol.Syst.https://doi.org/10.15252/msb.202311627* (2023).
72. Zhang, L., Zhang, J. & Nie, Q. DIRECT-NET: an efficient method to discover cis-regulatory elements and construct regulatory networks from single-cell multiomics data.*Sci. Adv.* 8, eabl7393 (2022).
73. Duren, Z., Chen, X., Jiang, R., Wang, Y. & Wong, W. H. Modeling gene regulation from paired expression and chromatin accessibility data.*Proc.Natl Acad.USA* **114**, E4914-E4923 (2017)。
74. Burdziak, C., Azizi, E., Prabhakaran, S. & Pe'er, D. A nonparametric multi-view model for estimating cell type-specific gene regulatory networks.Preprint at *arXiv* https://doi.org/ 10.48550/arXiv.1902.08138 (2019)。
75. Bachireddy, P. 等人.绘制T细胞在应答过程中的进化状态以及 对采用细胞疗法的耐药性.Cell Rep 37, 109992 (2021).*Cell Rep.* **37**, 109992 (2021)。
76. Kamimoto, K. et al.*Nature* **614**, 742-751 (2023).
这项工作从 **scRNA-seq** 和 **scATAC-seq** 数据中提出了一种新颖的 **GRN** 推断方法，同时还引入了硅学 **TF** 干扰策略。
77. Skok Gibbs, C. et al. 规模的高性能单细胞基因调控网络推断：Inferelator 3.0。*Bioinformatics* **38**, 2519-2528 (2022).
78. Fleck, J. S. et al. *自然*https://doi.org/10.1038/s41586-022-05279-8 (2022)。
79. Li, Z., Nagai, J. S., Kuppe, C., Kramann, R. & Costa, I. G. scMEGA: single-cell multi-omic enhancer-based gene regulatory network inference.*Bioinform.Adv.* **3**, vbad003 (2023).
80. Jiang, J. et al. IReNA：单细胞转录组 和染色质可及性图谱的综合调控网络分析。*iScience* **25**, 105359 (2022)。
81. 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。Preprint at *bioRxiv* https://doi.org/10.1101/ 2022.09.14.508036 (2022)。
82. Zhang, S. et al. Inference of cell type-specific gene regulatory networks on cell lineages from single cell omic datasets.*Nat.14, 3064 (2023).***14**, 3064 (2023).
83. 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间过程调控分析。*Genome Res.* **30**, 622-634 (2020)。
84. Xu, Q. et al. ANANSE: an enhancer network-based computational approach for predicting key transcription factors in cell fate determination.*Nucleic Acids Res.* **49**, 7966-7985 (2021).
85. Duren, Z. et al. sc-compReg 可利用单细胞数据比较不同条件下的基因调控网络。*Nat.12, 4763 (2021).***12**, 4763 (2021).
86. Kuppe, C. 等人. 人类心肌梗塞的空间多组图.*Nature* **608**、766-777 (2022).
87. Argelaguet, R. 等人. 利用单细胞 多组学解码小鼠胚胎的基因调控。Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.15.496239 (2022)。
88. Zeng, W. et al. DC3 是一种对 和单细胞基因组学数据进行解卷积和耦合聚类的方法。*Nat.Nat.***10**, 4613 (2019).
89. Liberzon, A. et al. The molecular signatures database (MSigDB) hallmark gene set collection.*Cell Syst.***1**, 417-425 (2015).
90. Anderson, A. G. et al. 单核多组学鉴定 ZEB1 和 MAFB 为阿尔茨海默病特异顺式调控元件的候选*调控*因子。*Cell Genomics* **3**, 100263 (2023).
91. Thompson, D., Regev, A. & Roy, S. 基因调控网络的比较分析： from network reconstruction to evolution.*Annu.Rev. Cell Dev.***31**, 399-428 (2015).
92. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data.*Genetics* **155**, 945-959 (2000).
93. Lou, S. et al. TopicNet: a framework for measuring transcriptional regulatory network change.*Bioinformatics* **36**, i474-i481 (2020).
94. Alvarez, M. J. et al. 利用 基于网络的蛋白质活性推断癌症体细胞突变的功能特征。*Nat.Genet.***48**, 838-847 (2016).
95. Badia-i-Mompel, P. et al. decoupleR: ensemble of computational methods to infer biological activities from omics data.*Bioinforma.Adv.* **2**, vbac016 (2022).
96. Subramanian, A. et al. 基因组富集分析：基于知识的全基因组表达谱解读方法。*Proc.Natl Acad.USA* **102**, 15545-15550 (2005).
97. Garcia-Alonso, L. 等人。转录因子活性增强癌症药物敏感性标记 。*Cancer Res.* **78**, 769-780 (2018)。
98. Walsh, L. A. et al. 综合系统生物学方法发现 TRIM25 是乳腺癌转移的关键决定因素。*Cell Rep.* **20**, 1623-1640 (2017).