Bioinformatics Advances, 2023, vbad003
https://doi. org/10. 1093/bioadv/vbad003Advance Access
Publication Date: 12 January 2023

Application Note

基因表达

scMEGA:基于单细胞多组增强子的基因调控网络推断

李志坚[1]， James S. Nagai [1], Christoph Kuppe [2,3], Rafael Kramann [2,3,4], Ivan G. Costa[1,*]

[1]亚琛工业大学医学院计算生物医学联合研究中心计算基因组学研究所，德国亚琛52062，[2]亚琛工业大学实验医学与系统生物学研究所，德国亚琛52062，[3]亚琛工业大学肾脏学与临床免疫学学部，德国亚琛52062，[4]伊拉斯谟医学中心内科、肾脏学与移植科，荷兰鹿特丹3042

*通信地址应注明收件人。副主编:Marieke
Lydia Kuijjer

2022年10月18日收;2022年12月8日修订;2023年1月5日编辑决定;2023年1月11日录用

摘要

摘要:单细胞多组学数据的日益可用性使得定量表征基因调控成为可能。我们在这里描述了scMEGA(基于单细胞多组学增强子的基因调控网络推断)，它能够对基因调控网络推断的多组学数据进行端到端分析，包括模式整合、轨迹分析、增强子-启动子关联、网络分析和可视化。这使得能够研究动态生物过程的复杂基因调控机制，如细胞分化和疾病驱动的细胞重塑。我们提供了人类心肌梗死中控制肌成纤维细胞活化的基因调控网络的案例研究。

可用性和实现:scMEGA是用R实现的，在MIT许可下发布，可从https://github.com/CostaLab/scMEGA获得。教程可从https://costalab.github.io/scMEGA获得。联系人:ivan.costa@rwth-aachen.de

补充信息:补充数据可在Bioinformatics Advances在线获取。

1 介绍

单细胞RNA测序(scRNA-seq)和ATAC-seq (scATAC-seq)技术通过捕获正交分子信息，为了解单细胞水平的基因调控提供了前所未有的机会(李志坚等人，2021;Zhu等人，2020)。将这两种分析方法应用于相同的生物样本，可以产生单细胞多组学数据，从而可以计算推断不同细胞系统的基因调控网络(grn)，例如苍蝇的大脑发育(Janssens等人，2022)和人类心肌功能(Kuppe等人，2022)。

然而，这种分析通常基于复杂的生物信息学管道，每个步骤都需要不同的工具，例如用于scRNA-seq分析和数据集成的Seurat (Stuart等人，2019)，用于scATAC-seq分析和轨迹推断的ArchR (Granja等人，2021)，用于TF活性估计的chromVAR (Schep等人，2017)和用于网络分析的igraph (Csardi和Nepusz, 2006)。目前，有三种计算工具(Pando, Fleck等，2022;CellOracle Kamimoto等人，2020;图r Kartha等人，2022)可用于基于单细胞多组学图谱的GRN推断。Pando专注于识别限于TF-TF相互作用的调节网络，而没有提供模态整合或轨迹分析的方法。CellOracle可独立分析ATAC-seq和RNA-seq数据。因此，它不能探索基因表达来描绘基因到峰值增强子的联系，也不能在单细胞水平上考虑TF活性评分来选择相关的转录因子。图r是目前最全面的方法之一，它包括多模态数据集成、峰-基因链接预测和轨迹推断等模块。然而，它不支持对Seurat对象的直接操作，并且提供很少的功能来从grn中探索信息。

我们在这里开发了scMEGA作为一个通用框架，通过将单细胞多组学谱作为输入，定量推断基于增强子的GRN。scMEGA基于对心肌梗死多模态单细胞数据分析的专业知识(Kuppe等人，2022)，能够对GRN推断的多组学数据进行端到端分析，包括模态集成、轨迹分析、增强子-启动子关联、网络分析和可视化(图1)。为此，它提供了新的功能，并结合了Seurat、ArchR、chromVAR和igraph的一些现有方法。它是作为R包实现的，并且与Seurat生态系统兼容。

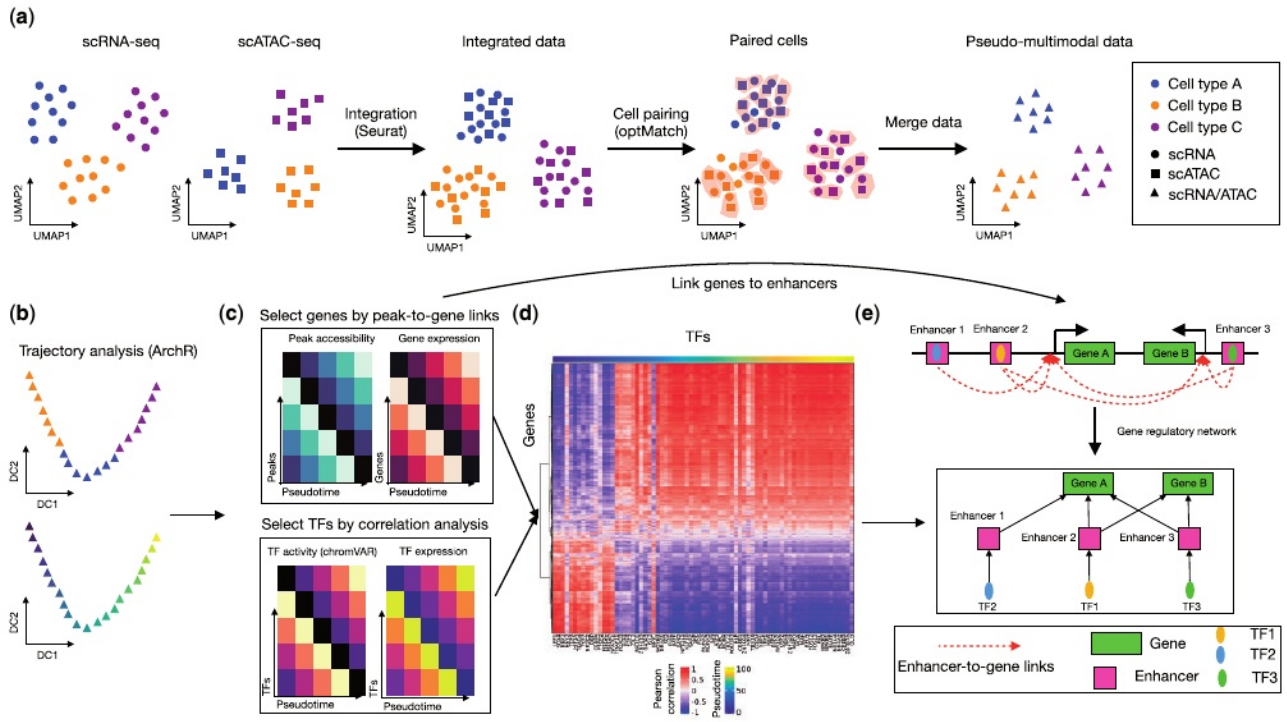

图1所示。scMEGA概述。(a)首先，scMEGA通过基于R软件包Seurat的模态集成和基于OptMatch方法的细胞配对，整合单细胞多组学图谱，获得配对数据集。(b)接下来，scMEGA利用ArchR推断伪时间轨迹来表征给定细胞类型的潜在动态过程。(c)然后过滤根据结合活性(chromVAR)与沿轨迹表达的相关性分析选择的tf，以及根据峰值可及性与沿轨迹基因表达的相关性分析选择的基因。(d) scMEGA根据所选择的tf和基因，通过估计所选择的TF与基因表达的相关性，生成定量的GRN。(e) scMEGA利用增强子-基因链接和基序匹配来寻找基于增强子的tf -基因相互作用。这些用于过滤先前定义的定量GRN，如(d)所示。

2 方法

scMEGA有三个主要步骤，即(i)多模态数据集成，(ii)候选tf和基因的识别和过滤，以及(iii) GRN组装和分析。

2.1 单细胞多模态数据集成

为了构建grn，在不同模式(RNA和ATAC)的细胞之间建立映射是至关重要的，这样就可以在单细胞水平上发现TF/基因的表达与基因/tf的活性和可及性之间的关联。为此，scMEGA首先使用Seurat实现的canonical correlation analysis (CCA)将细胞投影到共享的共嵌入空间(Stuart et al.， 2019)。这里的特征是来自scRNA-seq的基因表达和来自scATAC-seq数据的基因活性评分。如果存在批处理效果，则可以应用Harmony (Korsunsky et al.， 2019)方法进行批处理校正。接下来，scMEGA执行细胞配对，使用OptMatch配对获得scRNA-seq和scATAC-seq之间的一对一匹配(Kartha等，2022)。总之，这些步骤构建了一个伪多模态数据集和数据的低维表示(图1a)。在具有成对单元的单细胞多模态数据的情况下[例如SHARE-seq (Ma et al.， 2020)和10 Multiome]，可以跳过此集成步骤。在这里，用户只需要使用MOJITOO的CCA集成来创建单元的联合嵌入(Cheng et al.， 2022)。

2.2 候选tf和基因的鉴定

scMEGA接下来根据前一步的多模态或伪多模态数据识别候选tf和基因。首先，对于给定的感兴趣的细胞，使用R包ArchR实现的监督方法来推断表征潜在动态过程的伪时间轨迹(Granja等人，2021)(图1b)。在这里，用户需要指示根细胞和终末细胞，即通过标记基因对其进行表征。

然后，scMEGA根据染色质可及性谱估计每个细胞中每个TF的结合活性(图1c)。为了识别活性tf, scMEGA计算TF结合活性(用chromVAR估计;Schep et al.， 2017)和TF表达。高相关性表明TF高度表达，且基序比平均谱更容易获得(Janssens et al.， 2022)。这一步是至关重要的，因为tf结合活性本身不能区分具有相似基序的同一家族的tf。接下来，scMEGA沿着伪时间轨迹计算每个基因的表达变异，并挑选出前10%(这个截止值可以由用户调整)最可变的基因作为轨迹相关基因。接下来，scMEGA利用ArchR的功能，根据基因表达的相关性和单细胞水平的峰值可及性，将所选基因与峰值关联(Granja等人，2021)。在其他功能中，scMEGA提供了允许直接操作Seurat对象的功能。

2.3 GRN的构建和可视化

接下来，scMEGA通过估计所有选定tf的结合活性与所有选定基因表达的相关性，构建定量GRN，如上所述(图1d)。为了将TF与其靶基因连接起来，scMEGA首先将预测的峰-基因链接亚集以获得增强子-基因链接，其中增强子被定义为距离基因转录起始位点至少2k碱基对(bp)的峰(图1e)。然后考虑chromVAR预测的tf结合位点。如果一个基因与至少一个增强子相关，并且该TF与其中一个相关的增强子结合，则该基因仅被认为是TF的靶标，从而产生基于增强子的GRN。通过结合定量GRN和基于增强子的GRN，即我们只考虑两个网络中的tf基因调控，scMEGA产生最终的基于增强子的GRN (eGRN)。tf -基因的相互作用由它们的相关性来加权。

定向(从TF到基因)和加权(通过相关性测量)grn通过R包图建模为图(Csardi和Nepusz, 2006)。scMEGA通过探索布局算法(如Fruchterman - Reingold)提供eGRN可视化。这种布局允许通过绘制共享相似靶基因的TF来找到主要的调控模块。或者，用户可以使用焦点布局，它允许关于某些相关基因/TF的网络集中。scMEGA还通过计算所有TF或目标的页面排名指数(Page et al.， 1999)或间隔性得分(Freeman, 1978)来探索网络统计来描述重要的TF。之间性分数找到调节因子(TF)，它连接GRN的不同模块(Zaoli等人，2021)。页面排名检测重要的TF, TF直接或间接调节其他基因(Ghoshal和Baraba ' si, 2011)。scMEGA允许可视化TF靶点的基因表达，以了解空间坐标上的调控活性。

3 结果

3.1 对scMEGA的鲁棒性进行基准测试

多模态单细胞数据的整合是单细胞非配对数据scMEGA的重要步骤。为了测试这一步骤的影响，我们从人类健康外周血单核细胞中获得了使用10个 Multiome协议生成的单细胞多模态数据。我们回收了10 504个细胞，并鉴定了4种青少年细胞类型(补充图1a)。接下来，我们整合数据并进行细胞配对。我们观察到，只有少数真配对被正确恢复(n¼41)，这表明在单细胞水平上，细胞匹配确实是一项艰巨的任务。然而，大多数细胞与正确的细胞类型匹配(71.8%)，而不匹配通常代表相似的亚群，例如非经典和中间单核细胞(补充图1b)。这些数字与最近的基准研究(Lance et al.， 2022)具有竞争力。接下来，我们使用scRNA-seq和scATAC-seq数据之间的真实或预测对来预测CD4 T细胞的eGRN(补充图1c)。事实上，从OptMatch预测对推断出的75%以上的TF、83%的基因和60%的tf基因调控也得到了真实对的支持，这表明scMEGA可以恢复大多数相互作用(补充图1d-f)。

3.2 心肌梗死的案例研究

我们在此提供了一个案例研究，使用scMEGA推断GRN来研究心肌梗死后人类心脏的纤维生成(Kuppe等人，2022)。我们整合了snRNA-seq和snATAC-seq数据，鉴定出了4个成纤维细胞亚群(补充图2a)。对标记基因的检测表明，簇2高表达SCARA5，最近有报道称SCARA5是人肾脏肌成纤维细胞祖细胞的标记(Kuppe等，2021)(补充图2b)。集群1被POSTN、COL1A1和COL3A1标记，表明这些细胞是已分化的肌成纤维细胞。在此基础上，我们建立了从簇2到簇1的伪时间轨迹来研究肌成纤维细胞的分化过程(补充图2c)。

接下来，我们选择了79个候选TF和2207个基因作为推断GRN的输入(补充图2d和e)。TF的结合活性与基因表达之间的相关性揭示了两个主要的调控模块，每个模块对应于一个不同的成纤维细胞亚簇(补充图2f)。例如，我们发现NR3C2是SCARA5þ成纤维细胞(模块1 -成纤维细胞祖细胞)的调节因子，其结合活性、TF表达和靶基因表达沿轨迹下降(图2b)。对于肌成纤维细胞，我们检测到几种与纤维化相关的TF，如TEAD (Liu et al.， 2017)和RUNX家族基因。值得注意的是，

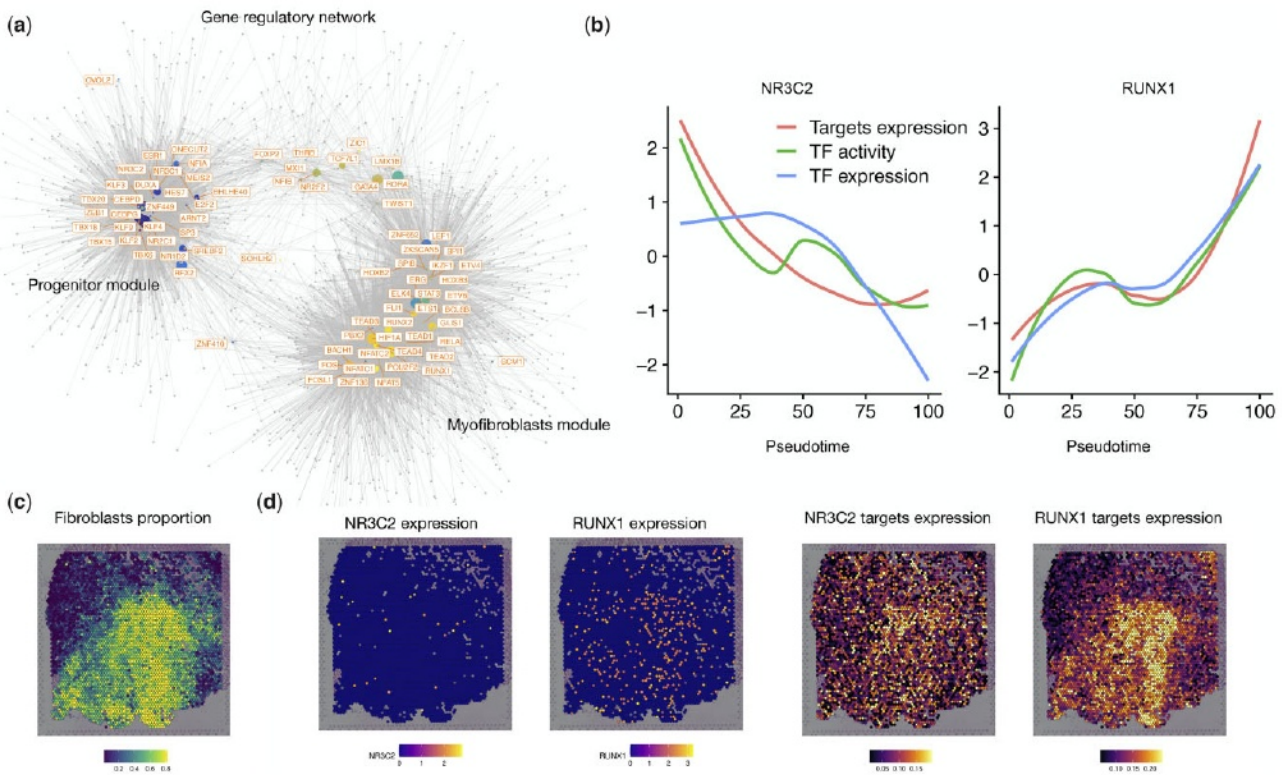

图2所示。推断GRN与肌成纤维细胞分化有关。(a)肌成纤维细胞分化的GRN可视化。每个节点代表一个TF(调节器)或基因(目标)。TF由chromVAR估计的TF具有最高活性评分的伪时间点着色。(b)线形图显示NR3C2和RUNX1沿肌成纤维细胞差异轨迹的TF结合活性、TF表达和靶表达。x轴表示假时间点，y轴表示z-score转换值。(c)心肌梗死后人心脏缺血区cell2定位估计成纤维细胞比例的空间分布可视化。(d)左:NR3C2和RUNX1基因的空间分布表达。右:NR3C2和RUNX1各靶点基因表达的空间分布

我们最近描述了RUNX1在肾脏(李志健等人，2021)和心脏(Kuppe等人，2022)纤维化中发挥重要作用。

可视化推断的网络和调节因子属性(在和页面排名分数之间)确定RUNX1是肌成纤维细胞分化过程中更重要的因素(补充图3a)。以RUNX1为中心的eGRN可视化突出了这样一个事实，即RUNX1被预测调节许多其他基因，包括其他与纤维化相关的转录因子TEAD2和TEAD3(补充图3b)。作为egrn允许的下游分析的另一个例子，我们在空间中检测了NR3C2和RUNX1靶基因的表达。由于这些TF在空间转录组学中的稀疏性和低表达值，我们无法检测到TF在空间上的清晰表达模式(图2c和d)。通过探索NR3C2和RUNX1的规则体(靶基因)，我们在纤维化反应的特定心脏区域观察到梯度和互排斥的空间表达，突出了scMEGA在稀疏空间转录组学数据中描绘TF规则体表达的能力。

4 结论

我们提出scMEGA来推断基于增强子的GRN使用单细胞多组学/多模态谱。scMEGA是建立在几个R包单细胞数据分析。它使用户能够执行端到端的GRN推断，并优先考虑重要的TF和基因，以进行实验验证，并使用规则体分析空间转录组学。我们举例使用scMEGA来研究心肌梗死后人类心脏中肌成纤维细胞活化的基因调控(Kuppe等，2022)。本文分析的数据集分别包含至少63000和20000个snRNA-seq和snATAC-seq细胞，这证明了scMEGA的可扩展性。此外，数据可以从不同的平台或协议生成，因为批处理效果将使用Harmony进行计算校正(Korsunsky et al.， 2019)。然而，轨迹分析假设细胞是分化或激活过程的一部分。此外，本文和其他人(Lance et al.， 2022)提出的单细胞匹配问题的基准测试表明，这是一个极其困难的问题。未来的工作包括实现额外的细胞匹配算法，如(Lance et al.， 2022)中报道的表现最好的方法。总之，我们认为scMEGA是理解单细胞多组学数据中各种生物过程的复杂基因调控机制的重要框架。

作者的贡献

李志健(概念化[主导]、数据策展[对等]、形式分析[对等]、方法论[对等]、资源[主导]、软件[主导]、写作原创稿[对等]、写作评论与编辑[对等])、James Shiniti Nagai(概念化[支持]、形式分析[支持]、软件[支持]、可视化[支持]、写作原创稿[支持])、Christoph Kuppe(概念化[支持]、数据策展[支持]、写作原创稿[支持])和Ivan G. Costa(概念化[平等]，资金获取[领导]，调查[领导]，方法论[领导]，项目管理[领导]，监督[领导]，写作原创稿[领导]，写作审查和编辑[领导])

软件和数据可用性

本文使用的数据可在:https://zenodo.org/record/6623588。本工作中提出的分析笔记可在 https://costalab.github.io/scMEGA/articles/myofibroblast-GRN.html找到。

资金

本项目由德国研究基金会(DFG)和德国教育和科学部(BMBF)资助的E: MED联盟纤维图谱资助。

利益冲突:未声明。

参考文献

Cheng,M. et al. (2022) MOJITOO: a fast and universal method for integration of multimodal single-cell data. Bioinformatics, 38, i282–i289.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. InterJournal Complex Syst., 1695, 1–9.

Fleck, J. S. et al. (2022) Inferring and perturbing cell fate regulomes in human cerebral organoids. Nature. https://doi.org/10.1038/s41586-022-05279-8.

Freeman,L.C. (1978) Centrality in social networks conceptual clarification. Soc. Networks, 1, 215–239.

Ghoshal,G. and Baraba´si,A.-L. (2011) Ranking stability and super-stable nodes in complex networks. Nat. Commun., 2, 1–7.

Granja,J.M. et al. (2021) Archr is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat. Genet., 53, 403–411.

Janssens,J. et al. (2022) Decoding gene regulation in the fly brain. Nature, 601, 630–636.

Kamimoto,K. et al. (2020) Celloracle: dissecting cell identity via network in-ference and in silico gene perturbation. bioRxiv. https://doi.org/10.1101/2020.02.17.947416.

Kartha,V.K. et al. (2022) Functional inference of gene regulation using single-cell multi-omics. Cell Genomics, 2, 100166.

Korsunsky,I. et al. (2019) Fast, sensitive and accurate integration of single-cell data with harmony. Nat. Methods, 16, 1289–1296.

Kuppe,C. et al. (2021) Decoding myofibroblast origins in human kidney fibro-sis. Nature, 589, 281–286.

Kuppe,C. et al. (2022) Spatial multi-omic map of human myocardial infarc-tion. Nature, 608, 766–777.

Lance,C. et al. (2022) Multimodal single cell data integration challenge: results and lessons learned. PMLR,176, 162–176.

Li,Z. et al. (2021) Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. Nat. Commun., 12, 1–14.

Liu,R. et al. (2017) Tead1 is required for maintaining adult cardiomyocyte function, and its loss results in lethal dilated cardiomyopathy. JCI Insight, 2, 1–15.

Ma,S. et al. (2020) Chromatin potential identified by shared single-cell profil-ing of RNA and chromatin. Cell, 183, 1103–1116.

Page,L. et al. (1999). The PageRank citation ranking: Bringing order to the web. Technical report. Stanford InfoLab.

Schep, A. N. et al. (2017) Chromvar: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat. Methods, 14, 975–978.

Stuart,T. et al. (2019) Comprehensive integration of single-cell data. Cell, 177, 1888–1902.

Zaoli,S. et al. (2021) Betweenness centrality for temporal multiplexes. Sci. Rep., 11, 1–9.

Zhu, C. et al. (2020) Single-cell multimodal omics: the power of many. Nat. Methods, 17, 11–14.