SCENIC:单细胞调节网络推断和聚类

Sara aibar1,2, Carmen Bravo González-Blas 1,2, Thomas moerman3, 4, v<e:1> n Anh Huynh-Thu5, Hana Imrichova1, 2, Gert hulselman1, 2, Florian rambow6,7, Jean-Christophe marine6,7, Pierre Geurts5, Jan aerts3, 4, Joost van den Oord8, Zeynep Kalender atak1, 2, Jasper Wouters1,2,8 & 斯坦Aerts1 2

我们提出了SCENIC，一种从单细胞RNA-seq数据中同时进行基因调控网络重建和细胞状态鉴定的计算方法(http://scenic.aertslab.org)。在一份来自肿瘤和大脑的单细胞数据纲要上，我们证明了*顺式调控*分析可以用来指导转录因子和细胞状态的鉴定。SCENIC为驱动细胞异质性的机制提供了重要的生物学见解。

细胞的转录状态来自于一个潜在的基因调控网络(GRN)，在这个网络中，有限数量的转录因子(tf)和辅因子相互调控，并调控它们的下游靶基因。单细胞转录组分析的最新进展为高分辨率鉴定转录状态和状态之间的转换(例如在分化过程中)提供了令人兴奋的机会1,2。针对单细胞RNA-seq进行优化的统计技术和生物信息学方法已经带来了新的生物学见解，但目前尚不清楚是否可以确定稳定细胞状态下的特异性和鲁棒性grn。这可能确实具有挑战性，因为在单细胞水平上，由于转录爆发和其他来源的基因表达随机变化，基因表达可能与TF输入的动态部分断开。已经开发了一些从单细胞RNA-seq数据推断共表达网络的方法5 - 7，但这些方法不使用调控序列分析来预测tf和靶基因之间的相互作用。

我们推断，将*顺式调控*序列与单细胞基因表达联系起来可以克服缺失和技术变异，从而优化细胞状态的发现和表征。为此，我们开发了单细胞调节网络推断和聚类(SCENIC)来绘制grn，然后通过评估每个细胞中grn的活性来识别稳定的细胞状态。SCENIC工作流包括三个步骤(图**1a**，**补充图1**，参见在线方法)。在第一步中，使用GENIE3鉴定与tf共表达的基因集(参考文献8)(补充**图1a)**。由于GENIE3模块仅基于共表达，因此它们可能包含许多假阳性和间接靶标。为了确定可能的直接结合靶点，每个共表达模块都使用RcisTarget进行*顺式调控*基序分析(补充**图1b**，参见在线方法)。只保留具有正确上游调控子显著基序富集的模块，并对其进行修剪以去除缺乏基序支持的间接目标。我们将这些处理过的模块称为调控子。

作为SCENIC的一部分，我们开发了AUCell算法来对每个细胞中每个调控子的活性进行评分(补充**图1c**和**2**，并参见在线方法)。对于给定的规则，比较细胞间的AUCell分数可以确定哪些细胞具有更高的子网络活动。得到的二进制活动矩阵降低了维数，这对下游分析很有用。例如，基于该矩阵的聚类基于调控子网络的共享活动来识别细胞类型和状态。由于规则子是作为一个整体进行评分的，而不是使用单个基因的表达，因此这种方法对辍学具有鲁棒性(补充**图3)**。

为了评估SCENIC的性能，我们将其应用于成年小鼠大脑中已知细胞类型的scRNA-seq数据集9(图**1b-e)**。该分析提供了151个规则-在1,046个初始共表达模块中-具有相应tf显著丰富的基序(占初始tf的7%)。对每个细胞的调节子活性进行评分，揭示了预期的细胞类型(图**1d,e)**以及每种细胞类型的潜在主调节因子列表(例如，**补充图4中的小胶质细胞网络)**。按细胞类型聚类(总灵敏度为0.88，特异性为0.99，调整后的Rand指数(ARI) > 0.80)比许多专用的单细胞聚类方法更准确10。

为了评估SCENIC的稳健性，我们重新分析了小鼠大脑数据:完整的数据集;随机选择100个细胞的样本来模拟小数据集;或者三分之一的测序读数来模拟低覆盖率数据。SCENIC识别出仅由少数细胞代表的细胞类型(例如，来自小胶质细胞、星形胶质细胞或中间神经元的2至6个细胞;**补充图5)**。此外，预测的tf与细胞类型的关联与先前建立的一致

1VIB Center for Brain & Disease Research, Laboratory of Computational Biology, Leuven, Belgium. 2KU Leuven, Department of Human Genetics, Leuven, Belgium. 3KU Leuven ESAT/STADIUS, VDA-lab, Leuven, Belgium. 4IMEC Smart Applications and Innovation Services, Leuven, Belgium. 5Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium. 6VIB Center for Cancer Biology, Laboratory for Molecular Cancer Biology, Leuven, Belgium. 7KU Leuven, Department of Oncology, Leuven, Belgium. 8KU Leuven, Department of Imaging and Pathology Translational Cell and Tissue Research, Leuven, Belgium. Correspondence should be addressed to S.A. (stein.aerts@kuleuven.vib.be).

Received 5 December 2016; accepted 7 September 2017; published online 9 OCTOBER 2017; doi:10.1038/nmeth.4463

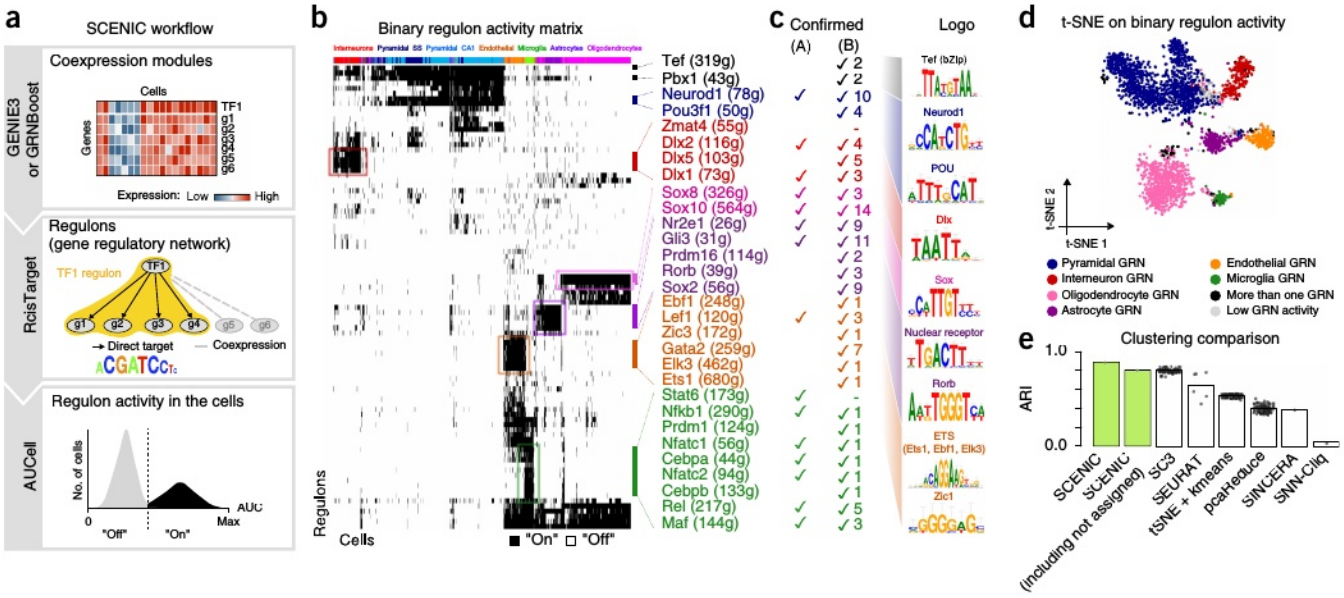

图1 | SCENIC工作流程及其在老鼠大脑中的应用。(a)在SCENIC工作流程中，首先使用GENIE3或GRNBoost推断tf和候选靶基因之间的共表达模块。然后，RcisTarget识别调控子结合基序在靶基因中显著富集的模块，并仅用直接靶标创建调控子。AUCell对每个细胞中每个调控子的活性进行评分，从而产生二值化的活性矩阵。细胞状态的预测是基于调控子网络的共享活动。(b)小鼠脑SCENIC结果。聚类标签与参考文献9中使用的标签相对应;主调节器的颜色与它们控制的细胞类型相匹配。(c)经文献(A)证实或具有小鼠基因组信息学(B)的脑表型的tf;其相应的富集dna结合基序显示(Logo)。(d)二元调控子活性矩阵上的t-SNE。每个细胞被分配最活跃GRN的颜色。(e)不同聚类方法在该数据集上的准确率。

角色(图**1c**)，并且这种准确性优于标准分析管道(补充**图3e**)。

为了验证为小鼠中间神经元鉴定的DLX1/2网络，我们分析了人类大脑的单核RNA-seq数据集11(补充**图6**)。在人类数据上，SCENIC还鉴定了一组由DLX1/2强烈驱动的中间神经元，它们具有与小鼠相同的识别基序，并鉴定了一组包括DLX1本身在内的保守靶标(图**2a,b**)。接下来，我们将这种跨物种分析扩展到其他细胞类型12。与基于归一化表达的标准聚类相反，后者产生了强大的物种驱动聚类(补充**图7**)，SCENIC分析有效地按细胞类型对细胞进行分组(图**2c**)。这表明网络活动的评分是鲁棒的，可以用来克服批量或技术影响(补充**图3d**)。

我们还应用SCENIC在来自少突胶质细胞瘤13(来自6个肿瘤的4043个细胞)和黑色素瘤14(来自14个病变的1252个细胞)的scRNA-seq数据集中鉴定复杂的细胞状态。由于肿瘤特异性突变和复杂的基因组畸变，癌细胞状态的鉴定比正常细胞状态的鉴定更具挑战性15。标准的聚类方法根据细胞的起源肿瘤对其进行分组(图**3a,b**)，但SCENIC显示了不同的情况。对于少突胶质细胞瘤，在肿瘤中鉴定出三种癌细胞状态(图3c-**e**)，每种状态都由预期的tfs驱动，包括SOX10/4/8、OLIG1/2和ASCL1用于少突胶质细胞样状态;SOX9、NFIB和AP-1为星形细胞样状态;E2F和FOXM1代表循环细胞。

此外，将扩散图应用于二进制SCENIC矩阵(补充**图8**)，重建了从茎样分支到少突胶质细胞样分支和星形胶质细胞样分支的分化轨迹。请注意，与正常少突胶质细胞分化相比，这条路径代表了不同的“轨迹”(参见**补充图9**，对5069个胶质细胞进行了SCENIC分析)。我们在黑色素瘤数据上观察到类似的肿瘤效应校正，其中SCENIC识别出肿瘤中的细胞群(补充**图10**)，包括由少突胶质细胞瘤中类似的tf驱动的一组循环细胞(例如，E2F1/2/8和MYBL2;**图3f-h**及**补充图10**)。与专用的批效应去除方法(如Combat16和Limma17)不同，SCENIC通过使用生物驱动特征自动去除肿瘤效应，**这些方法需要先验地指定批效应的来源(补充图**11)。

黑色素瘤细胞大致分为两组，一组对应于MITF高状态(典型的增殖状态，MITF和STAT/IRF是关键的调节因子)，另一组对应于MITFlow状态，WNT5A、LOXL2和zeb1的表达上调，这些都是侵袭状态的已知标记(补充**图10e,f**)。SCENIC鉴定出两种处于MITFlow状态的新tf, NFATC2(114个预测靶基因)和NFIB(15个预测靶基因)。NFATC2是JNK/MAPK通路中的转录抑制因子，参与黑色素瘤去分化和免疫逃逸18。另一方面，NFIB与毛囊和黑素细胞干细胞的干细胞行为有关，并在小细胞肺癌的转移进展中起重要作用。

为了进一步探索NFATC2和NFIB在MITFlow状态中的潜在作用，我们对25例肿瘤进展不同的黑色素瘤标本进行了免疫组织化学处理。我们发现NFIB和NFATC2在前哨淋巴结的表达最高。这与ZEB1表达共定位，这表明这些标记物的表达与前列腺癌之间存在关系

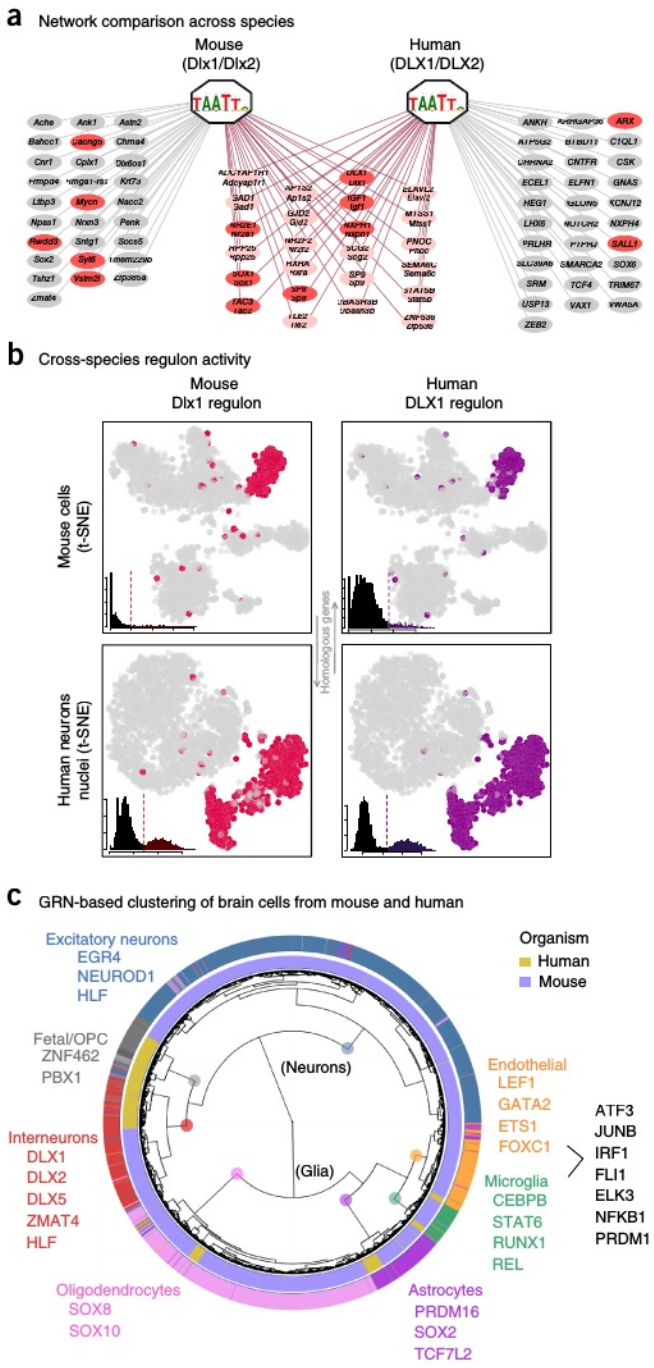

图2 |神经网络和细胞类型的跨物种比较。(a)从小鼠和人脑scRNA-seq数据推断的DLX1/2调控子。红色的基因与GeneMANIA的Dlx1/2相关。(b)人和小鼠Dlx1/2调控在小鼠和人单细胞数据上的互反活性。在每个SCENIC t-SNE图中，细胞根据相应的二进制调控子活性被着色。插图说明了规则的AUCell分数分布。(c)基于GRN活性对人和小鼠脑scRNA-seq数据进行联合聚类。彩色TF名称对应于人类和小鼠SCENIC运行中确定的规则。

最早的转移事件(图**3i**和**补充图12**)。当我们在A375(一种NFATC2和NFIB高表达的黑色素瘤细胞系)中使用siRNA敲低NFATC2时(补充**图13**)，我们发现NFATC2调控基因显著上调(见在线方法)。这与之前确定的NFATC2作为抑制因子的作用是一致的。此外，参与细胞粘附和细胞外基质调控的基因以及先前发表的代表黑色素瘤侵袭状态的几个基因特征也被上调(补充**表1**)，这表明NFATC2可能确实在疾病进展中发挥了重要作用。作为对黑色素瘤调控子的第二次验证，我们使用ChIP-seq数据确认了MITF和STAT的预测靶标(图**3j**)。

随着单细胞数据集规模的增加，我们建议采用两种互补的方法来扩展网络推理。第一种方法是从子采样数据集推断GRN，并在AUCell评分步骤中包含所有单元格。我们在具有来自小鼠视网膜的40,000多个单细胞的数据集上说明了这种方法(补充**图14**)。第二种方法旨在使用更有效的机器学习和大数据处理解决方案。我们在Scala和Apache Spark上实现了GENIE3的新变种GRNBoost，用梯度提升取代了随机森林回归。这种实现大大减少了推断GRN所需的时间(补充**图15**)，并将为在非常大的数据集上进行网络推断铺平道路，例如即将推出的人类细胞图谱22。

SCENIC是一种普遍适用于scRNA-seq数据分析的方法，它利用tf*和顺式调控*序列来指导细胞状态的发现。我们的研究结果表明，grn构成了识别细胞状态的强大指南，并且scRNA-seq数据非常适合于追踪基因调控程序，其中tf的特定组合驱动细胞类型特异性转录组。

方法

方法，包括数据可用性声明和任何相关的加入码和参考文献，可在论文的在线版本中获得。

注意:任何补充信息和源数据文件都可以在论文的在线版本中获得。

致谢

本工作由研究基金会-弗兰德斯(FWO;授予S. Aerts G.0640.13和G.0791.14;鲁汶大学特别研究基金(PF/10/016和OT/13/103资助S. Aerts)，抗癌基金会(2012-F2, 2016-070 和 2015-143 资 助 S. Aerts) 和 ERC consolidated Grant (724226_顺- control资助S. Aerts)。S. Aibar得到了鲁汶大学PDM博士后奖学金的支持。zk.a.和J.W.由Kom op Tegen Kanker提供博士后奖学金;v.a.h.t。得到比利时frs - fnrs的支持;艾滋病毒是由科学技术创新机构(IWT)的博士奖学金支持的。T.M.和J.A.的资助由Symbiosys和IMEC HI^2数据科学提供。资助者在研究设计、数据收集和分析、决定发表或准备手稿方面没有任何作用。T.M.要感谢J. Simm关于梯度提升的有用意见和建议。

作者的贡献

S. Aerts和S. Aibar构思了这项研究;S. Aibar在v.a.h.t的帮助下实现了SCENIC和相关软件包。GENIE3为P.G.， RcisTarget为G.H.;S. Aibar和cbg - b。在zk.a.和艾滋病毒的帮助下分析数据;T.M.和J.A.实现了GRNBoost;J.W.进行了免疫组化和敲除实验;f.r.、j.c.m.和J.v.d.O.提供了试剂，并帮助解释黑色素瘤的分析;S. Aibar、J.W.和S. Aerts撰写了手稿。

竞争的经济利益

作者声明没有经济利益竞争。

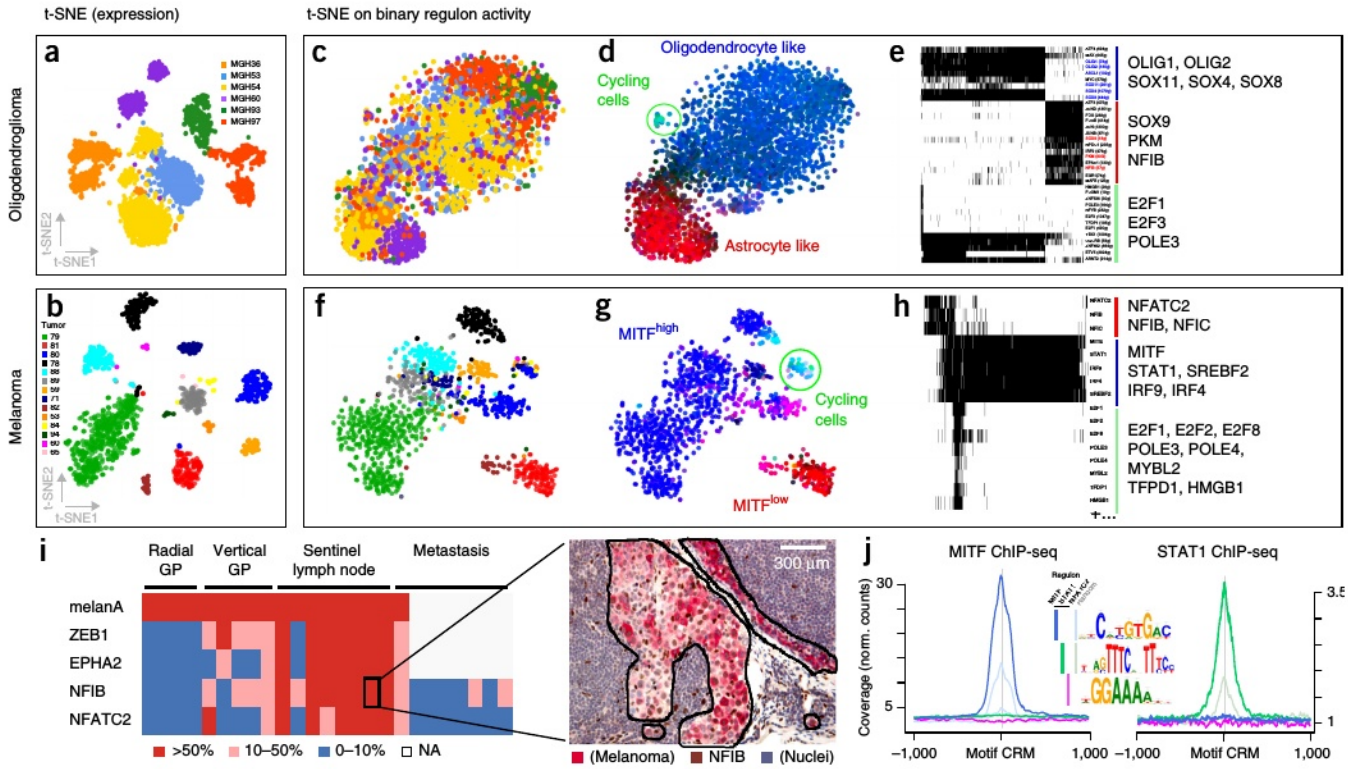

图3 | SCENIC克服了肿瘤效应，揭示了肿瘤中相关的细胞状态和grn。(a,b)基于表达矩阵的t-SNE图，以肿瘤原点着色。(c,d和**f,g)**应用SCENIC后基于二元活性矩阵(e,h)的t-SNE图。在**d**和**g中**，细胞通过GRN活性着色。(i)使用NFATC2、NFIB、ZEB1和EPHA2抗体对25例人黑色素瘤进行免疫组化(IHC)。热图显示了给定样本中每种标记物呈阳性的细胞百分比。右图为前哨淋巴结NFIB IHC的典型例子(其他图像见**补充图13)**。NA，不适用。(j) MITF和STAT1 ChIP-seq信号在预测目标区域和随机选择的以MITF/STAT基序出现为对照的基因组区域的聚集图。

1. Linnarsson, S. & Teichmann, S.A. *Genome Biol.* **17**, 97 (2016).
2. Wagner, A., Regev, A. & Yosef, N. *Nat. Biotechnol.* **34**, 1145–1160 (2016).
3. Stegle, O., Teichmann, S.A. & Marioni, J.C. *Nat. Rev. Genet.* **16**, 133–145 (2015).
4. Raj, A. & van Oudenaarden, A. *Cell* **135**, 216–226 (2008).
5. Moignard, V. *et al. Nat. Biotechnol.* **33**, 269–276 (2015).
6. Pina, C. *et al. Cell Rep.* **11**, 1503–1510 (2015).
7. Guo, M., Wang, H., Potter, S.S., Whitsett, J.A. & Xu, Y. *PLoS Comput. Biol.* **11**, e1004575 (2015).
8. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L. & Geurts, P. *PLoS One* **5**, e12776 (2010).
9. Zeisel, A. *et al. Science* **347**, 1138–1142 (2015).
10. Kiselev, V.Y. *et al. Nat. Methods* **14**, 483–486 (2017).
11. Lake, B.B. *et al. Science* **352**, 1586–1590 (2016).
12. Darmanis, S. *et al. Proc. Natl. Acad. Sci. USA* **112**, 7285–7290 (2015).
13. Tirosh, I. *et al. Nature* **539**, 309–313 (2016).
14. Tirosh, I. *et al. Science* **352**, 189–196 (2016).
15. Alizadeh, A.A. *et al. Nat. Med.* **21**, 846–853 (2015).
16. Johnson, W.E., Li, C. & Rabinovic, A. *Biostatistics* **8**, 118–127 (2007).
17. Ritchie, M.E. *et al. Nucleic Acids Res.* **43**, e47 (2015).
18. Perotti, V. *et al. Oncogene* **35**, 2862–2872 (2016).
19. Chang, C.-Y. *et al. Nature* **495**, 98–102 (2013).
20. Denny, S.K. *et al. Cell* **166**, 328–342 (2016).
21. Müller, M.R. & Rao, A. *Nat. Rev. Immunol.* **10**, 645–656 (2010).
22. Regev, A. *et al. bioRxiv* Preprint at: http://www.biorxiv.org/content/early/2017/05/08/121202 (2017).

网上的方法

SCENIC工作流程。SCENIC是基于三个新的R/bioconductor软件包的工作流程:(i) GENIE3，用于基于共表达识别潜在的TF靶标;(ii) RcisTarget，进行TF-motif富集分析并确定直接靶标(regular -lons);(iii) AUCell，对单个细胞上的调控子(或其他基因集)的活性进行评分。我们还提供了在Spark23上实现的GRNBoost，作为在更大数据集上构建共表达网络的可扩展替代方案(步骤1，取代GENIE3)。

三个R/bioconductor包和GRNBoost包括详细的教程，以方便他们在自动化SCENIC管道中使用，以及独立的工具。这些工具、SCENIC代码和教程的链接可在http://scenic.aertslab.org上找到。

GENIE3。GENIE3(参考文献8)是从基因表达数据推断基因调控网络的方法。简而言之，它训练随机森林模型来预测数据集中每个基因的表达，并使用tf的表达作为输入。然后使用不同的模型来获得tf的权重，测量它们各自与预测每个目标基因表达的相关性。最高的权重可以转化为tf靶调控链接8。由于GENIE3使用随机森林回归，它具有允许TF与其候选目标之间复杂(例如非线性)共表达关系的附加价值。GENIE3在Python, Matlab和R中可用。为了允许包含在SCENIC工作流中，我们优化了以前的GENIE3 R实现。这个新实现的核心现在是用C语言编写的(这使得它的速度提高了几个数量级)，它需要更低的内存，并且它支持并行执行。在DREAM4和DREAM5挑战赛中，GENIE3是表现最好的网络推理方法24。新包在DREAM挑战中提供了与以前现有实现类似的结果，但速度有所提高。这个对比可以在以下网站上找到:http://www.montefiore.ulg.ac.be/~huynh-thu/GENIE3.html。

GENIE3的输入是一个表达式矩阵。首选的表达值是基因汇总计数(可能使用也可能不使用唯一的分子标识符UMIs25)。其他度量，如计数或每百万记录(TPM)和FPKM/RPKM也被接受为输入。然而，请注意，第一个网络推理步骤是基于共表达式的，一些作者建议避免样本内归一化(即TPM)，因为它们可能会导致人为的协变26。为了评估输入矩阵的归一化在多大程度上影响SCENIC的输出，我们还在库大小归一化后的Zeisel等人9的数据集上运行了SCENIC(使用scran27的标准管道，它执行簇内大小因子归一化)。结果具有高度可比性，无论是在产生的细胞簇或细胞类型(从原始UMI计数或归一化计数获得的细胞类型之间的ARI: 0.90，与作者的细胞类型相比，归一化计数的ARI: 0.87)和识别组的tf(**图1b中突出显示的30个规则中的26个**)。此外，在本项目过程中，我们将GENIE3应用于多个数据集，其中一些具有UMI计数(例如，小鼠脑和少突胶质细胞)，另一些具有TPM计数(例如，人脑和黑色素瘤)，两个单元都提供了可靠的结果。

GENIE3的输出是一个包含基因、潜在调节因子及其“重要性度量”(IM)的表格，IM表示TF(输入基因)在预测目标中的权重。我们探索了几种确定阈值的方法(例如，查看使用RcisTarget进行修剪后的排名、分布和输出)，并最终选择为每个TF构建多个潜在目标的基因集:(i)设置几个IM阈值(IM > 0.001和IM > 0.005)，(ii)为每个TF获取最高IM的50个目标，(iii)仅保留每个靶基因的前5个，10个和50个TF(然后，按TF分割)。在所有这些情况下，只考虑了IM > 0.001的链接。此外，将每个基因集分成正相关和负相关靶标(即TF与潜在靶标之间的Spearman相关性)，以分离可能被激活和被抑制的靶标。最后，只保留至少有20个基因的基因集(TF共表达模块)用于下一步。

GRNBoost。GRNBoost基于与GENIE3相同的概念:纯粹从基因表达矩阵推断每个靶基因的调控因子。然而，GRNBoost使用XGBoost库29中的梯度提升机(GBM)28实现来实现这一点。GBM是一种集成学习算法，它使用boostin30作为一种策略，将多个弱学习器(如浅树)组合成一个强学习器。这与随机森林形成对比，随机森林是GENIE3使用的方法，它使用bagging (bootstrap aggregation)进行模型平均以提高回归精度。GRNBoost使用梯度增强树桩(深度为1的回归树)31作为基础学习器。GRNBoost的主要贡献是将这种多元回归方法转换为基于Apache Spark23的Map/Reduce32框架。在GRNBoost中，核心数据条目是一个由基因名称和基因表达值向量组成的元组。GRNBoost首先使用Spark RDD将基因表达向量划分到计算集群中可用的节点上。随后，它构建一个包含所有候选调控基因的表达值的预测因子矩阵。使用Spark广播变量，预测矩阵被广播到不同的计算分区。在框架的映射阶段，GRNBoost遍历基因元组(表达向量)，并使用预测器矩阵以表达向量作为各自的训练标签来训练XGBoost回归模型。从训练好的模型中，提取出调控因子-目标关系的优势，并作为一组网络边发出。在约简阶段，所有的边集合被组合成最终的调节网络。

GRNBoost和GENIE3在2个Intel Xeon E2696 V4 cpu、总共44个物理核或88个线程和128 GB 2133Ghz ECC内存的工作站上的性能进行了比较。大型数据集和因此产生的大型预测器矩阵导致网络推断成为内存约束而不是CPU约束。为了将所需的内存量舒适地放入可用的128 GB内存中，我们将分区数量减少到11个，因此同时运行的预测器矩阵最多只有11个。但是，我们将每个XGBoost回归的可用线程数增加到8个，有效地使用了工作站中的所有可用线程(88个)。GRNBoost是用Scala编程语言编写的，可以作为软件库使用，也可以从命令行作为Spark作业提交。

RcisTarget。RcisTarget是一个新的R/Bioconductor实现i-cisTarget和iRegulon的基序富集框架。

RcisTarget识别丰富的tf结合基序和候选转录因子的基因列表。简而言之，RcisTarget基于两个步骤。首先，它选择在基因集中基因的转录起始位点(transcription start site, TSS)周围显著过度代表的DNA基序。这是通过在数据库上应用基于恢复的方法实现的，该数据库包含每个基序的全基因组跨物种排名。注释到相应TF并获得归一化富集评分(NES) > 3.0的基序被保留。接下来，对于每个基序和基因集，RcisTarget预测候选目标基因(即，在基因集中排名在前沿之上的基因)。该方法基于Aerts等人33*描述的方法，该方法也在i-cisTarget (web界面)34和iRegulon (Cytoscape插件)35中实现。因此，当使用相同的参数和数据库时，RcisTarget提供与i-cisTarget或iRegulon相同的结果，与Janky等人的其他tfbs*富集工具进行基准测试35。关于该方法及其在R中的实现的更多细节在包文档中给出。为了构建最终的调控，我们合并了每个TF模块的预测靶基因，这些靶基因显示了给定TF的任何基序的富集。为了检测抑制，理论上可以采用与负相关TF模块相同的方法。然而，在我们分析的数据集中，这些模块的数量较少，并且显示出非常低的基序富集。出于这个原因，我们最终决定从工作流中排除直接抑制的检测，只继续研究正相关的目标。本文分析使用的数据库是来自iRegulon(基于基因的motif排名)的人类和小鼠的“18k motif collection”。对于每个物种，我们使用了两个基因基序排序(TSS周围10 kb或TSS上游500 bp)，这决定了跨TSS周围的搜索空间。

AUCell。AUCell是一种新方法，允许研究人员在单细胞RNA-seq数据中识别具有活性基因调控网络的细胞。AUCell的输入是一个基因集，输出是每个细胞中的基因集“活性”。在SCENIC中，这些基因集就是调控子，由tf和它们假定的靶标组成。AUCell将调节子的富集计算为恢复曲线下的面积(AUC)，该面积横跨特定细胞中所有基因的排名，其中基因根据其表达值进行排名。因此，该方法独立于基因表达单位和归一化过程。此外，由于细胞是单独评估的，因此它可以很容易地应用于更大的数据集(例如，如果需要，对表达矩阵进行子集)。简而言之，评分方法基于恢复分析，其中*x轴*(补充**图1c)**是基于表达水平的所有基因的排名(具有相同表达值的基因，例如“0”，随机排序);而*y轴*则是从输入集中恢复的基因数量。然后，AUCell使用AUC来计算输入基因集的关键子集是否在每个细胞的排名顶部富集。这样，AUC就代表了在签名中表达基因的比例，以及它们与细胞内其他基因的相对表达值。这一步的输出是一个矩阵，其中包含每个细胞中每个基因集的AUC分数。我们直接使用AUC分数(跨规则)作为连续值来聚类单个细胞，或者我们使用每个规则的AUC分数的截止值来生成二进制矩阵。这些截止点是自动确定的，或者通过检查AUC分数的分布来手动调整。**补充图2a中提供了一些AUC分布的示例**。**补充图2b,c**显示了使用先前发表的神经元和胶质基因签名对AUCell的验证。包中包含的教程还包括该方法每个步骤的实际解释和含义。

基于基因调控网络的细胞聚类。细胞调节子活性总结在一个矩阵中，其中列代表细胞，行代表调节子。在二进制规则-活动矩阵中，与给定细胞中的活动规则对应的矩阵坐标将包含“1”，否则包含“0”。等效矩阵包含每个细胞调控子的连续AUC值，通常称为AUC活性矩阵。对任意一个调节子活性矩阵进行聚类，可以揭示出在细胞子集中周期性活跃的调节子组(联合起来，形成一个网络)。二元活性矩阵倾向于突出跨细胞的高阶相似性(因此高度减少了批处理效应和技术偏差);另一方面，AUC矩阵允许研究人员观察到更细微的变化。对于可视化，我们主要使用t-SNEs (Rtsne package36，我们总是测试几个困惑值和距离度量/ pc数量的一致性)，以及具有分层聚类的热图(尽管热图图形具有选择规则，但t-SNEs总是在整个矩阵上运行)。在教程中，我们还包含了几个选项来探索结果。例如，如何检测最可能的稳定状态(t-SNE中的高密度区域)，并帮助识别可能与检测状态相关的关键调节因子、已知细胞特性(基于数据集注释)和GO术语(调控簇中基因的GO富集分析)。

SCENIC运行在不同的数据集上。使用作者提供的表达矩阵(从GEO或作者网站下载)在所有数据集上运行SCENIC，仅包括通过质量控制的细胞，以及GENIE3的默认基因过滤(在所有这些数据集中产生12,000-15,000个基因)。标准SCENIC工作流程在所有数据集上运行(出版时的软件包版本可作为**补充软件获得**，更新版本将发布在http://scenic.aertslab.org)。对数据集和每种分析的任何特性的更详细描述可在**补充说明1中获得**。在这里，我们提供了数据集的简要描述:

小鼠皮层和海马体。幼鼠(21 ~ 31 d) 3005个脑细胞单细胞rna测序。它包含海马和体感觉皮层的主要细胞类型，即神经元(锥体兴奋性神经元和中间神经元)、胶质细胞(星形胶质细胞、少突胶质细胞、小胶质细胞)和内皮细胞。表达式矩阵单位:UMI计数。(Zeisel等9,GSE60361)

人类的神经元。对正常人脑3083个神经元细胞进行单核rna测序(从一位51岁女性的尸体中提取，来自6个不同的Brodmann区)。表达式矩阵单位:TPM。(Lake等11)

人类的大脑。466个成人和胎儿大脑细胞的scrna测序。这些胎儿样本取自4个不同的个体，时间为妊娠后16 - 18周。成年人的大脑

doi:10.1038/nmeth.4463

在治疗难治性癫痫和海马硬化症的颞叶切除手术中，从8名不同患者(21-63岁)的健康颞叶组织中提取样本。表达矩阵单位:记录的CPM。(Darmanis等*12*,GSE67835)

鼠标少突胶质细胞。来自少突胶质细胞谱系的5069个细胞的scrna测序数据。细胞从几种不同的小鼠品系中获得，并从小鼠幼年和成年中枢神经系统的前后轴和背腹轴的十个不同区域分离，包括白质和灰质。表达式矩阵单位:UMI计数。(Marques *et al.37*, GSE75330)

少突神经胶质瘤。6例未经治疗的IDH1或IDH2突变和1p/19q编码的II级少突胶质细胞瘤4347个细胞的scrna测序表达谱仅使用肿瘤细胞进行分析(由作者根据CNV谱选择)。表达式矩阵单位，log2(TPM + 1). (Tirosh等，*13*,GSE70630)

黑素瘤。对来自14种不同肿瘤的1252个黑色素瘤细胞进行scrna测序。这些仅包括作者根据其CNV谱标记为恶性的细胞。表达式矩阵单位:log2(TPM/10 + 1). (Tirosh *et al.14*, GSE72056)

老鼠的视网膜。通过Drop-seq获得小鼠视网膜44,808个细胞的scrna测序数据(出生后14 d)。表达矩阵单位:log((细胞中每个基因的UMI计数/细胞中总UMI计数)×10,000) + 1) (Macosko *et al.38*, GSE63472)

胚胎小鼠大脑。Chronium Megacell演示数据集包含来自两只E18小鼠(品系:C57BL/6)的皮质、海马和室下区1,306,127个细胞。(10 x基因组学)

基因筛选。对于运行GENIE3的基因过滤，我们基于基因计数总数和检测到的细胞数量应用了软过滤器。第一个过滤器，即每个基因的总读取数，旨在去除最可能不可靠且只提供噪声的基因。具体数值取决于数据集;对于本文中使用的那些，我们将阈值设置为，例如，3个UMI计数(略高于非零值的中位数)乘以数据集中细胞数量的1%(例如，在小鼠大脑中:3个UMI计数×30(细胞的1%)=每个基因最少90个计数)。第二个过滤器，即检测到基因的细胞数量(例如，> 0 UMI，或>1 log2(TPM))，是去除仅在一个或极少数细胞中表达的基因(如果它们碰巧在给定细胞中一致，它们将获得大量重量)。在工作流程中，我们建议将第二次过滤设置为低于要检测的最小细胞群。例如，由于小胶质细胞约占数据集中总细胞的3%，我们使用的检测阈值至少为细胞的1%。

跨物种网络比较。SCENIC对用于GRN比较的三个数据集分别独立运行:Zeisel等*人9*(小鼠脑细胞)，Lake等人*11*(人类神经元核)和Darmanis等人*12*(人类脑细胞)。为了跨物种比较网络，使用Biomart(通过Biomart R package39)将人类调控子中的基因转换为同源小鼠基因，反之亦然(将小鼠调控子转换为人类基因)。在**图2a中**，红色突出显示的基因也与GeneMANIA40中的Dlx1/2(蛋白-蛋白相互作用、遗传相互作用、共表达或文献提及)存在关联。

对于跨物种细胞聚类(图**2c)**，将小鼠表达矩阵中的基因转化为同源的人类基因，并逐行与Darmanis等*12*表达矩阵合并(仅保留两个矩阵中可用的基因)。Darmanis等*12*数据集中的259个人类调控子和小鼠调控子的人类同源物在该合并矩阵上进行评估，得到包含410个调控子的二元调控子活性。基于二元活性矩阵，使用Ward的分层聚类和Spearman的距离对细胞进行聚类。相反的方法(将表达矩阵转换为小鼠基因以评估小鼠调控)也获得了类似的结果。为了提供一种仅基于表达的替代方法(补充**图7)**，我们还生成了一个合并的表达矩阵。由于合并的数据集使用不同的测量单位(人类的CPM和小鼠的UMI)，因此在合并之前，每个矩阵都经过基因的Z-score归一化。

方法比较。我们进行了不同的评估和基准比较，每个评估SCENIC的不同方面(例如，细胞类型鉴定、TF鉴定、联合发现效应校正)。如何进行这些比较的详细描述可在**补充说明1中获得**。在这里，我们提供一个简短的总结:

细胞集群。为了确定基于基因调控网络活性的聚类是否与真实细胞类型相匹配，我们将基于调控子活性矩阵的聚类与相应出版物中提供的细胞标签进行了比较。为了将SCENIC性能与其他方法进行比较，我们重用了SC3 publication10中提供的基准，该基准为六种聚类方法在小鼠大脑数据集上提供了调整后的Rand指数(ARI)。

tf基序发现。SCENIC鉴定的TF的验证主要是通过在文献中确认它们在给定细胞类型中的作用来完成的(例如，**图1e)**。然而，我们还将SCENIC与另一种鉴定可能调节细胞状态的TF的方法进行了比较——对集群之间差异表达的基因(即基因标记或细胞类型的标记)应用TF基序富集分析。

批处理效应校正。将SCENIC对少突胶质细胞瘤数据集(完整二进制调控活性矩阵的聚类)的结果与combat16,41和limma17,42进行比较，校正了“原患者”作为批处理效应的来源。

骑自行车的细胞。基于amiGO和cycleBase 1.0和2.0中46个与有丝分裂细胞周期相关的基因集的一致上调，预测了循环细胞。然后，我们比较了不同聚类方法识别这些细胞的能力(敏感性和特异性)。由于大多数方法都提供了多个聚类作为输出，为了比较它们的结果，对于每种方法，我们选择了CC细胞数量最多的聚类。

黑色素瘤活组织检查的免疫组织化学。用黑素ana、EPHA2、ZEB1、NFATC2和NFIB抗体免疫组化对福尔马林固定的、嵌套的黑色素瘤样本进行检测。样本包括9例原发性黑色素瘤(4例为放射状生长期，5例为垂直生长期)、8例含黑素瘤前哨淋巴结和8例黑色素瘤转移灶的活检。**补充说明1中详细描述了如何进行免疫组织化学，以及使用的抗体**。

黑色素瘤细胞培养中*NFATC2*的下调。通过比较来自COSMIC Cancer cell lines Project43的59个黑色素瘤细胞系，基于*NFATC2*、*NFIB* (Supplementary **Fig. 13)**和*SOX10*的表达，选择A375细胞系作为MITFlow状态的代表。用*NFATC2* siRNA在A375细胞中敲低*NFATC2*，敲低72 h后提取总RNA。对最终文库进行汇总，并在NextSeq 500和HiSeq 4000 (Illumina)的组合上进行测序。RNA-seq读数被映射到基因组(hg19)进行上游分析。方法的详细描述，包括细胞系来源、*NFATC2*的敲除、RNA-seq协议和生物信息学分析，可在**补充说明1中获得**。

代码的可用性。与SCENIC相关的软件包和教程的更新链接可在http://scenic.aertslab.org;the上获得，发布时的软件包版本作为**补充软件提供**。

数据可用性声明。*NFATC2*敲低RNA-seq数据已存储在NCBI的基因表达Omnibus44中，可通过GEO登录号GSE99466访问。**图1-3**的源数据文件可在线获取。A **Life Sciences Reporting Summary**可在此获取。

23. Zaharia, M. *et al.* In *Proc. of the 9th USENIX Conference on Networked Systems Design and Implementation* 2–2 (USENIX Association, 2012).
24. Marbach, D. *et al. Nat. Methods* **9**, 796–804 (2012).
25. Islam, S. *et al. Nat. Methods* **11**, 163–166 (2014).
26. Crow, M., Paul, A., Ballouz, S., Huang, Z.J. & Gillis, J. *Genome Biol.* **17**, 101 (2016).
27. Lun, A.T.L., McCarthy, D.J. & Marioni, J.C. *F1000Res.* **5**, 2122 (2016).
28. Friedman, J.H. *Ann. Stat.* **29**, 1189–1232 (2001).
29. Chen, T. & Guestrin, C. In *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, 2016).
30. Freund, Y. & Schapire, R.E. *Jinko Chino Gakkaishi* **14**, 771–780 (1999).
31. S☒awek, J. & Arodz´, T. *BMC Syst. Biol.* **7**, 106 (2013).
32. Dean, J. & Ghemawat, S. *Commun. ACM* **51**, 107–113 (2008).
33. Aerts, S. *et al. PLoS Biol.* **8**, e1000435 (2010).
34. Herrmann, C., Van de Sande, B., Potier, D. & Aerts, S. *Nucleic Acids Res.* **40**, e114 (2012).
35. Janky, R. *et al. PLoS Comput. Biol.* **10**, e1003731 (2014).
36. Krijthe, J. Rtsne: t-distributed stochastic neighbor embedding using Barnes-Hut implementation https://github.com/jkrijthe/Rtsne (2015).
37. Marques, S. *et al. Science* **352**, 1326–1329 (2016).
38. Macosko, E.Z. *et al. Cell* **161**, 1202–1214 (2015).
39. Durinck, S. *et al. Bioinformatics* **21**, 3439–3440 (2005).
40. Warde-Farley, D. *et al. Nucleic Acids Res.* **38**, W214–W220 (2010).
41. Leek, J. sva: Surrogate Variable Analysis. R package version 3.24.4 (2017).
42. Smyth, G. limma: Linear models for microarray data. (2015).
43. Forbes, S.A. *et al. Nucleic Acids Res.* **45**, D777–D783 (2017).
44. Edgar, R., Domrachev, M. & Lash, A.E. *Nucleic Acids Res.* **30**, 207–210 (2002).

doi:10.1038/nmeth.4463

nature research

通讯作者:Stein Aerts

☐ 初稿修订版本 ☐ ☒ 最后提交

生命科学报告摘要

自然研究希望提高我们发表的工作的可重复性。此表格旨在与所有被接受的生命科学论文一起发表，并为报告的一致性和透明度提供结构。每一篇生命科学投稿都将使用此表格;有些清单项目可能不适用于个别稿件，但为了清晰起见，所有领域都必须填写。

有关本表格中所包含的要点的进一步信息，请参见报告生命科学研究。有关自然研究政策的更多信息，包括我们的数据可用性政策，请参阅作者和审稿人以及编辑政策清单。

”实验设计

1. 样本大小

描述样本量是如何确定的。

SCENIC分析:我们展示了SCENIC对8个数据集的应用。这些数据集被选择用于涵盖不同的案例研究:明确定义的“静态”细胞类型(小鼠大脑)、发育过程(小鼠少突胶质细胞，该数据集是在多个发育数据集中选择的，以便与之前的分析进行比较)、跨物种比较(2倍人类大脑)和癌症(黑色素瘤和少突胶质细胞瘤)。所有这些数据集的范围都在1k-5k细胞之间。此外，我们还纳入了一个稀疏数据集(使用Drop-seq的49k小鼠视网膜细胞)和一个更大的数据集(Megacell演示)。

IHC:我们根据人类标本的可用性，每个类别选择了~5个黑色素瘤样本，并对4个放射状生长期原发性黑色素瘤、5个垂直生长期原发性黑色素瘤、8个前哨淋巴结转移瘤和8个完全转移瘤进行了免疫组织化学染色(结果见手稿)。

NFATC2敲除的RNA-seq:我们对A375细胞进行了NFATC2敲除，每个类别有一个重复，因为我们使用了全基因组，差异表达基因的排名列表(见下文)。NFATC2敲除基因的两个重复，测序覆盖率较低(约200万高质量读数)，可靠地再现了原始发现(数据未在手稿中显示)。

2. 数据除外

描述任何数据排除。

SCENIC分析:没有数据被排除在分析之外。由于分析是在公共数据集上进行的，我们使用了作者选择的单元格。任何进一步的选择都在方法中进行了描述(例如少突胶质细胞瘤:CNV，小鼠视网膜亚采样和小鼠大脑亚采样)。

IHC和NFATC2敲除的RNA-seq:没有数据被排除在分析之外。

3. 复制

描述实验结果是否符合可靠地重现。

SCENIC分析:SCENIC在所有分析数据集中可靠地识别出预期的细胞类型(加上一些新的细胞类型)。计算重复(子抽样)也提供了可重复的结果(补充图5和15)。

IHC:4例放射状生长期原发性黑色素瘤，5例垂直生长期原发性黑色素瘤，8例前哨淋巴结转移瘤和8例完全转移瘤进行染色(结果见稿件)。

RNA-seq on NFATC2敲低:NFATC2敲低的两个重复，测序覆盖率较低(约200万高质量读数)，可靠地再现了原始发现(数据未在手稿中显示)。

4. 随机化

描述样本/生物体/参与者的情况

SCENIC分析:不相关。每个分析都是独立的。

Nature Methods: doi:10.1038/nmeth.4463

allocated into experimental groups.

IHC:不相关。分组根据临床诊断确定。

NFATC2敲除的RNA-seq on NFATC2:不相关。选择黑色素瘤细胞系是基于cosmos细胞系面板中NFATC2的高水平。

5. 致盲

描述调查人员在数据收集和/或分析过程中是否对分组分配不知情。

SCENIC分析:仅对表达矩阵进行分析;
不考虑细胞类型或任何其他表型信息
由数据集作者提供。仅在分析结束时，为
验证时，将细胞型/表型数据与提供的聚类进行比较
由SCENIC提供。
IHC:病理学家对染色结果进行评分，染色结果是不透明的。

NFATC2敲除的RNA-seq:不相关。

注:所有涉及动物和/或人类研究参与者的研究必须披露是否使用了盲法和随机化。

6. 统计参数-已确认

对于所有使用统计方法的图形和表格，请确认在相关图例中(如果需要额外的空间，也可以在方法部分中)存在以下项目。

- ☐ ☒ ———— 每个实验组/条件的确切样本量(n)，以离散数字和测量单位(动物、窝、培养物等)给出
- ☐ ☒ 对如何收集样本的描述，注意测量是取自不同的样本还是相同的样本
  重复测量样本
- ☐ ☒ 表明每个实验被重复多少次的声明
- ☐ ☒ 所使用的统计检验，以及它们是单侧还是双侧(注:只有常见的检验应该只通过名称来描述;更多的复杂的技术应该在方法部分描述)
- ☐ ☒ 对任何假设或修正的描述，如对多重比较的调整
- ☐ ☒ 测试结果(如*P*值)尽可能以精确值给出，并注明置信区间
- ☒ ☐ 对统计数据的清晰描~~述，包括集中趋~~势(如中位数、平均值)和~~变异(~~如标准差、四分位数间距)
- ☒ ☐ 明确定义的误差条

进一步的资源和指导,请参阅生物学家统计数据的web集合。

”软件

有关计算机代码可用性的政策信息

7. 软件

描述用于分析本研究数据的软件。

SCENIC算法:本文提出了一种新的算法SCENIC，实现所需的三个新的r包(GENIE3、RcisTarget和AUCell)，以及GRNboost作为GENIE3的可扩展替代方案。所有这些都在方法中进行了描述，并且它们在R中的实现可以在Github中获得。R包也作为补充代码提供，新版本的链接将保存在作者网站(http://scenic.aertslab.org)上。

SCENIC分析:本文中介绍的分析是使用软件包的开发版本运行的。这些版本作为补充代码提供(仅接口在不同版本之间发生了变化):SCENIC 0.1.5(2017年7月17日)、AUCell 0.99.5(2017年6月7日)、RcisTarget 0.99.0(2017年6月7日)和GENIE3 0.99.3。使用RcisTarget的18k-motif数据库(Human: RcisTarget.hg19. motifdatabases_0.99.0, Mouse: RcisTarget.mm9. motifdatabases_0.99.0)进行分析。

公共数据集的补充分析:- R版本3.3.2和Bioconductor版本3.4对应的软件包。-基准:Homer(版本4.9), Seurat(版本1). -基因集富集分析:GSEA(版本2.0) GeneMANIA(访问时间:2016年10月)，amigo, cycleBase(1.0和2.0)。

Nature Methods: doi:10.1038/nmeth.4463

NFATC2上的RNAseq敲除:
- fastq-mcf(作为应用程序的一部分);版本1.1.2-686):使用包含常见Illumina适配器的列表的默认参数;从原始读取中修剪适配器序列。

-来自Babraham Bioinformatics的FastQC:用于修剪reads的质量控制。- STAR(版本2.5.1b-foss-2014a):将reads映射到人类refseq hg19基因组。

- SAMtools (version1.4-foss-2014a): 仅过滤-q4质量的reads - HTSEQ-count (version 0.6.1p1): 计算每个基因的reads数-来自Bioconductor的DESeq2 (version 1.14.1)用于R-studio:获取差异表达基因列表，根据向上或向下调节的Log2FC进行排名。

- GOrilla (cbl-gorilla.cs.technion.ac.il/):一个在线工具，用于识别富集的基因本体类别，基于全基因组，差异表达基因的排名列表。- GSEA(2.0版)，使用GSEAPreranked功能:一个Java桌面应用程序来评估全基因组中基因集的潜在富集，差异表达基因的排名列表。IHC:不相关。

对于使用自定义算法或软件的手稿，这些算法或软件是论文的核心，但尚未在已发表的文献中描述，必须根据编辑和审稿人的要求提供软件。我们强烈鼓励在社区存储库(例如GitHub)中存放代码。*Nature Methods*提供用于发布的算法和软件的指南提供了有关此主题的进一步信息。

”材料和试剂

关于材料可用性的政策信息

8. 材料的可用性

说明是否对独特材料的可用性有限制，或者这些材料是否只能由营利性公司分发。	N/A

Nature Methods: doi:10.1038/nmeth.4463

9. 抗体

描述所使用的抗体以及它们如何在研究系统中被验证使用(即测定和物种)。

IHC在徕卡BOND-MAX自动免疫染色机上进行(徕卡微系统)。抗原回收使用基于柠檬酸盐的(键表位检索溶液1,pH 6.0;徕卡)或基于edta的(键表位检索溶液2,pH 9.0;根据制造商的说明(见下文)使用徕卡缓冲。

用于染色的一抗:
-兔多克隆抗nfib (Sigma-Aldrich;HPA003956;pH6.0):抗体为通过Human Protein Atlas进行了广泛的选择性/特异性验证用于免疫组织化学染色。内部染色条件为在人体胰腺切片上验证。
-兔单克隆抗nfatc2(细胞信号技术);# 5861;pH6.0):抗体由供应商验证其选择性/特异性及其用于免疫组织化学染色。内部对染色条件进行验证人体淋巴结切片。
-兔多克隆抗zeb1 (Santa Cruz;sc - 25388;pH9.0):抗体验证由供应商进行选择性/特异性，并将其用于免疫组织化学对人黑色素瘤的染色由Caramel及其同事(Caramel et al.， Cancer Cell, 2013)。内部染色条件进行了验证人类黑色素瘤切片。
-兔单克隆抗epha2(细胞信号技术);# 6997;pH9.0):抗体由供应商验证其选择性/特异性及其用于免疫组织化学染色。内部对染色条件进行验证人类黑色素瘤切片。
-小鼠单克隆抗黑色素瘤抗体(DAKO;IR633;pH9.0):抗体通过供应商的选择性/特异性及其用于免疫组织化学染色。内部的染色条件在人类黑色素瘤上得到了验证部分。

用于染色的二抗:-用于棕色可视化:Bond Polymer Refine检测试剂盒(徕卡)-用于红色/粉色可视化:Bond Polymer Refine red Detection(徕卡)

10. 真核细胞系a.说明所使用的每种真核细胞系的来源。b.描述所使用的细胞系鉴定方法。

A375黑色素瘤细胞系由合作者(Lionel Larue教授，居里研究所，巴黎)提供。

根据ATCC验证A375细胞中存在BRAF纯合子c.1799T>A (p.V600E)、CDKN2A纯合子c.181G>T (p.E61*)和CDKN2A纯合子c.205G>T (p.E69*)三个突变。

c.报告是否对细胞系进行了支原体污染检测。

A375细胞系定期检测支原质污染。结果为阴性。

d.如果使用的任何细胞系在ICLAC维护的常见错误鉴定细胞系数据库中列出，请提供其使用的科学依据。

N/A

“动物和人类研究参与者

涉及动物研究的政策信息;在报告动物研究时，请遵循ARRIVE指南

11. 研究动物描述

提供研究中使用的动物和/或动物源性材料的详细信息。

N/A

Nature Methods: doi:10.1038/nmeth.4463

12。人类研究参与者描述

描述人类研究参与者的协变量相关群体特征。

对于IHC，我们在不知道患者年龄或性别的情况下为每个临床病理类别选择了黑色素瘤样本(~5)，但基于人类标本的可用性，并对4个放射状生长期原发性黑色素瘤，5个垂直生长期原发性黑色素瘤，8个前哨淋巴结转移瘤和8个完全转移瘤进行了免疫组织化学染色(结果见文稿)。

IHC实验已获得比利时鲁汶大学大学医院(BioMel;比利时参考编号B322201524395)，并通过鲁汶生物库(参考编号S57760)。

RNAseq实验也得到了鲁汶大学医院(ML10660;比利时参考编号B322201421305)，并由鲁汶生物库(参考编号S56777)提供。

整个研究符合《世界医学协会赫尔辛基宣言》。

Nature Methods: doi:10.1038/nmeth.4463