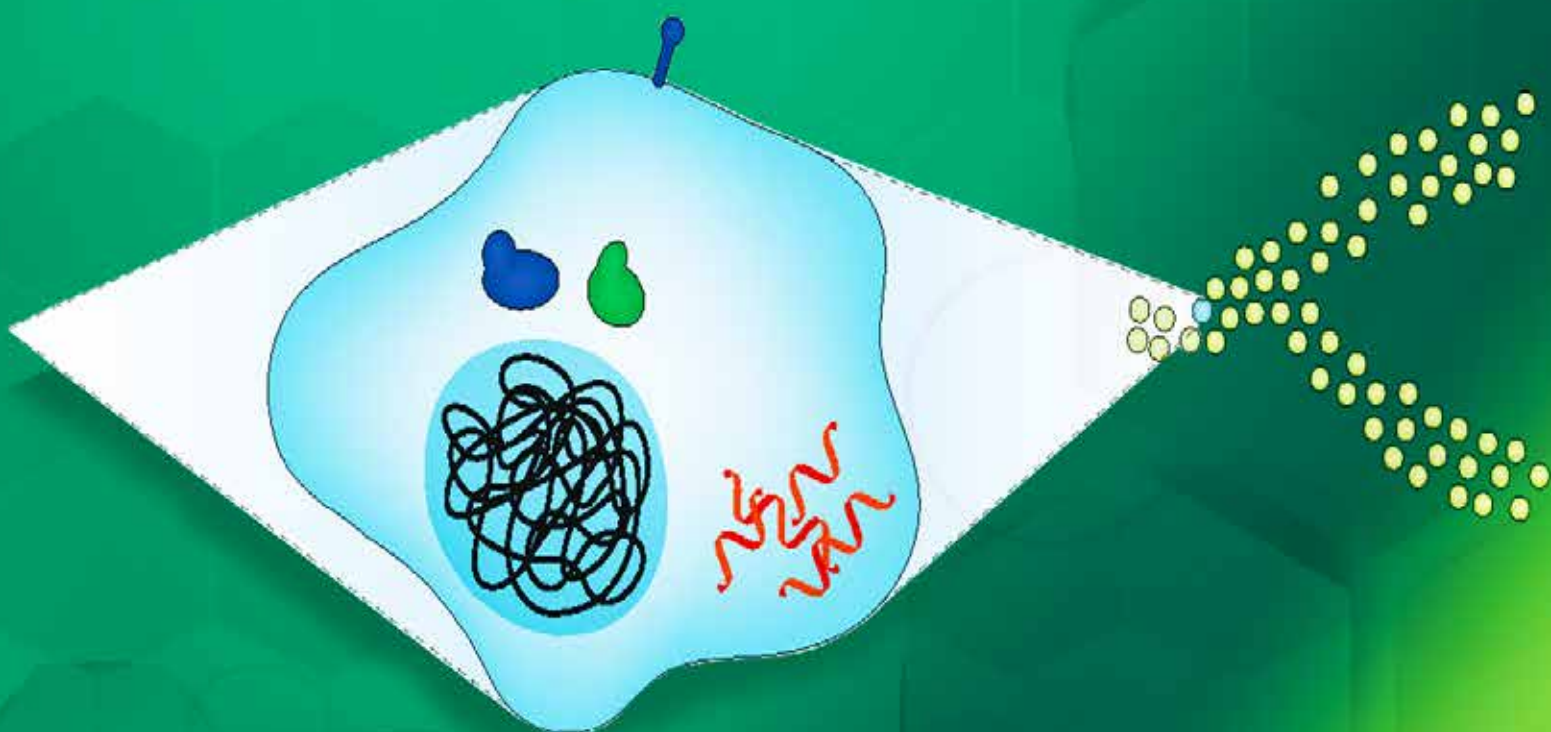


生命奥秘

总119期 / 2019/10

LIFEOMICS

单细胞多元数据 整合分析



无奇不有
生命世界
解读生命
走进科学

目录 CONTENTS

专题：单细胞多元数据整合分析

前言	02
----	----

一、单细胞多重检测	07
-----------	----

1. 在开展破坏性试验之前先收集细胞计量信息（cytometric information）	07
2. 分离细胞组份	09
3. 将细胞分析数据提炼成分子模型	10
4. 透过scRNA-seq数据发掘更多信息	12
5. 分析多元数据	13

二、整合不同单细胞试验获得的数据	15
------------------	----

1. 整合单细胞RNA测序数据	15
2. 将不同单细胞RNA测序数据集里的细胞进行分类	18
3. 多元单细胞数据的整合	20

三、将测序结果与空间数据相结合	22
-----------------	----

四、前景展望	25
--------	----

专题

单细胞多元 数据整合分析

前言

不断成熟的单细胞RNA测序技术（single-cell RNA sequencing, scRNA-seq）已经与其它革命性新技术一起，从遗传（genetic）、表观遗传（epigenetic）、空间（spatial）、蛋白质组（proteomic）和谱系（lineage）等各个方面，为我们揭开了单细胞的面纱。虽然还存在计算方面的问题，但这些科技进展也为我们提供了前所未有的机会，让我们能够利用多种不同的技术，对不同种类的细胞进行联合研究。这种综合分析（Integrated analysis）能够发现不同细胞之间的关系，全面了解不同细胞状态的意义，将各个单独的研究和技术获得的数据综合起来进行整体分析。本文，我们将探讨近年来在单细胞层面上综合分析研究的最新进展，重点关注综合分析在基因表达数据方面的应用。

最近，在分子生物学技术、微流体技术和纳米技术等方面取得的进展，使得单细胞测序成为了可能（图1）。最初的技术主要集中在某一个方面的检测上，比如DNA测序、RNA表达检测或染色质组装检测等。虽然这些技术也让我们对细胞的多样性（cellular diversity）和发育（development）等方面获得了革命性的新发现，但由于技术的限制，所以只能开展彼此割裂的研究，不过这种“各自为政”的研究方法还是限制了我们更深入地了解和认识单细胞内各种生物分子之间的关系。了解这些生物分子之间的相互作用，是深入了解细胞状态的关键，不过这也正是单细胞分析领域的难点。此外，随着所获数据的不断增加，以及数据获

取更加方便，则更需要新的算法，来帮助我们对这些数据进行标准化（normalization），以便对不同样本之间的研究结果进行比较，哪怕有明显的批次效应（batch effects），或者个体间的差异（interindividual variation）。

单细胞RNA测序技术是目前应用得最普遍的一种单细胞测序技术，它具有很高的灵敏度，具有多重处理能力，或称综合条码式处理能力。与此同时，我们在单细胞遗传、表观遗传、蛋白质组学、单细胞基因组测序、染色质组装、DNA甲基化、细胞表面蛋白、小RNA、组蛋白修饰和染色体构象等方面的研究工作也都取得了不错的进展。而且，最近取得的这些成果也帮助我们更加准确地掌握了单细胞的空间信息和谱系信息（图1、表1）。

一个理想化的试验流程应该对细胞的方方面面都进行全面的观察和研究，这包括一套完整的记录，涉及到分子状态、空间位置与环境的相互作用等。虽然这已经超出了现有技术的检测范围，但是利用多种技术和综合算法，我们还是可以尽可能地达成这一目标的。接下来，我们就将介绍转录组学、基因组学、表观遗传组学和蛋白质组学目前在单细胞层面上的研究水平，还将重点介绍多种数据之间的整合情况。我们将单细胞RNA测序数据与其它数据结合在一起进行分析，因为这也是目前最流行的试验方法。另外，我们还将介绍如何综合利用这些研究手段，对同一细胞开展科学研究。最后，我们将深入探讨这些检测技术的优劣，以及未来的应用潜力和发展方向。

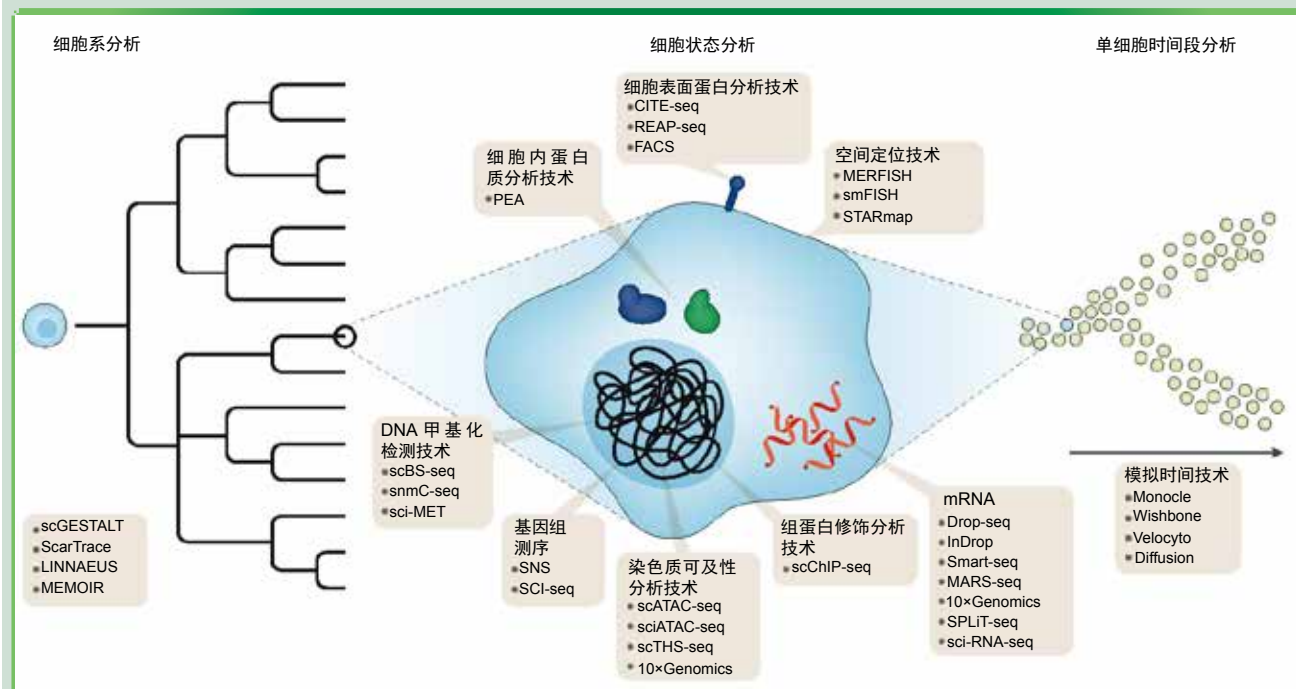


图1 单细胞多重综合分析。上图综合介绍了目前常用的单细胞综合分析技术。使用各种单细胞检测技术，可以在单细胞层面了解多种不同的细胞学参数。检测的范围大致可以分为细胞现有状态、细胞系信息，以及一段时间内的细胞变化情况。scGESTALT (single-cell genome editing of synthetic target arrays for lineage tracing)：利用人工合成的靶标芯片进行单细胞基因组编辑，用于进行细胞谱系追踪；ScarTrace：疤痕追踪技术，该技术能够添加荧光蛋白转基因的串联拷贝，从而能在CRISPR-Cas9基因编辑的转录过程中有效识别所遗留的“疤痕”；LINNAEUS (lineage tracing by nuclease-activated editing of ubiquitous sequences)：利用泛素序列核酸酶编辑技术进行细胞谱系追踪；MEMOIR (memory by engineered mutagenesis with optical in situ readout)：人工诱变光学原位检测记忆；scBS-seq (single-cell bisulfite sequencing)：单细胞亚硫酸盐测序；snmC-seq (single-nucleus methylcytosine sequencing)：单细胞核甲基胞嘧啶测序；sci-MET (single-cell combinatorial indexing for methylation analysis)：单细胞综合标记甲基化分析；PEA (proximity extension assay)：邻近延伸分析；SNS (single-nucleus sequencing)：单细胞核测序；SCI-seq (single-cell combinatorial indexed sequencing)：单细胞综合标记测序；CITE-seq (cellular indexing of transcriptomes and epitopes by sequencing)：细胞转录组及抗原表位标记测序；REAP-seq (RNA expression and protein sequencing assay)：RNA表达及蛋白质测序；FACS (fluorescence-activated cell sorting)：荧光细胞分选；scATAC-seq (single-cell assay for transposase-accessible chromatin using sequencing)：以测序技术为平台的单细胞转座酶可及染色质检测；sciATAC-seq (single-cell combinatorial indexing assay for transposase-accessible chromatin using sequencing)：利用测序技术进行单细胞转座子可及染色质组合标记分析；scTHS-seq (single-cell transposome hypersensitivity site sequencing)：单细胞转座子高敏位点测序；MERFISH (multiplexed error robust fluorescence in situ hybridization)：多重

低误差荧光原位杂交技术；smFISH（single-molecule fluorescence in situ hybridization）：单分子荧光原位杂交技术；STARmap（spatially resolved transcript amplicon readout mapping）：转录扩增子空间定位技术；scChIP-seq（single-cell chromatin immunoprecipitation followed by sequencing）：单细胞染色质免疫沉淀测序技术；Drop-seq：由哈佛医学院Steven McCarroll领导的团队开发，是基于使用微流控技术快速分析数千个单个细胞的策略，通过将它们封装在微滴中进行并行分析。利用这种方法，McCarroll等人已经分析了来自小鼠眼睛视网膜的44,800多个细胞的基因活性，鉴定出39种不同类型的视网膜细胞。他们的下一步目标是构建大脑的整体结构，以便对参与正常神经发育的基因有一个清晰的了解。InDrop：由哈佛大学的系统生物学教授Marc Kirschner领导开发，与Drop-seq有些类似，但不完全相同。研究小组利用这种方法来分析小鼠的数千个胚胎干细胞和分化细胞，以便更好地了解干细胞分化。Drop-seq的微珠库中有1600万个条形码，而inDrop方法只产生约15万个条形码，这意味着它每次运行处理的细胞数较少。不过，当研究人员试图分析极少量的组织样本时，inDrop可能更有优势，因其捕获细胞的比例高于Drop-seq。Smart-seq2：Smart-Seq（Switching mechanism at 5' end of the RNA transcript）是一项具有里程碑意义的技术，2012年由美国和瑞典的科学家共同开发。作为一种单细胞测序方案，它在转录本的序列覆盖度上有所改善。基因组的完整覆盖实现了选择性转录本异构体和SNV的检测。Smart-Seq2一年后发表在《Nature Methods》上，对最初的Smart-Seq方案做了一些改进。新方案使用锁核酸（LNA）、更高浓度的MgCl₂以及甜菜碱。它不再需要纯化步骤，可大大提高产量。MARS-seq（massively parallel RNA single-cell sequencing）：单细胞大规模平行RNA测序；SPLiT-seq（split-pool ligation-based transcriptome sequencing）：分裂池基于配位的转录组测序技术；sci-RNA-seq（single-cell combinatorial indexing RNA sequencing）：单细胞综合标记RNA测序。

表1 目前单细胞单参数检测及多参数检测方法一览表

检测对象	方法	检测范围	检测细胞通量
单参数检测方法			
mRNA测序技术	Drop-seq	转录组	1,000–10,000
	InDrop	转录组	1,000–10,000
	10× Genomics	转录组	1,000–10,000
	Smart-seq2	转录组	100–300
	MARS-seq	转录组	100–300
	CEL-seq	转录组	100–300
	SPLiT-seq	转录组	≥50,000
	sci-RNA-seq	转录组	≥50,000

续下表

续上表

基因组测序技术	SNS	基因组	10–100
	SCI-seq	基因组	10,000–20,000
染色质可及性检测技术	scATAC-seq	基因组	1,000–2,000
	sciATAC-seq	基因组	10,000–20,000
	scTHS-seq	基因组	10,000–20,000
DNA甲基化检测技术	scBS-seq	基因组	5–20
	snmC-seq	基因组	1,000–5,000
	sci-MET	基因组	1,000–5,000
	scRRBS	简化基因组	1–10
组蛋白修饰检测技术	scChIP-seq	基因组+一种修饰	1,000–10,000
染色体构象检测技术	scHi-C-seq	基因组	1–10
多参数检测方法			
组蛋白修饰+空间检测技术	NA	一个位点+一种修饰	10–100
mRNA+细胞系检测技术	scGESTALT	转录组	1,000–10,000
	ScarTrace	转录组	1,000–10,000
	LINNAEUS	转录组	1,000–10,000
细胞系+空间检测技术	MEMOIR	NA	10–100
mRNA+空间检测技术	osmFISH	10–50 RNAs	1,000–5,000
	STARmap	20–1,000 RNAs	100–30,000
	MERFISH	100–1,000 RNAs	100–40,000
	seqFish	125–250 RNAs	100–20,000
mRNA+细胞表面蛋白检测技术	CITE-seq	转录组+蛋白质	1,000–10,000
	REAP-seq	转录组+蛋白质	1,000–10,000
mRNA+染色质可及性检测技术	sci-CAR	转录组+基因组	1,000–20,000
mRNA+DNA甲基化检测技术	scM&T-seq	基因组	50–100
mRNA+基因组DNA检测技术	G&T-seq	转录组+基因组	50–200
mRNA+细胞内蛋白检测技术	NA	96 mRNAs+38 proteins	50–100
		82 mRNAs+75 proteins	50–200
DNA甲基化+染色质可及性检测技术	scNOME-seq	基因组	10–20

续下表

表1 附注

CEL-seq (cell expression by linear amplification and sequencing): 细胞线性扩增表达及测序; CITE-seq (cellular indexing of transcriptomes and epitopes by sequencing): 细胞转录组标记及表位测序; G&T-seq (genome and transcriptome sequencing): 基因组及转录组测序; LINNAEUS (lineage tracing by nuclease-activated editing of ubiquitous sequences): 利用泛素序列核酸酶编辑技术进行细胞谱系追踪; MARS-seq (massively parallel RNA single-cell sequencing): 单细胞大规模平行RNA测序; MEMOIR (memory by engineered mutagenesis with optical in situ readout): 人工诱变光学原位检测记忆; MERFISH (multiplexed error-robust fluorescence in situ hybridization): 多重低误差荧光原位杂交技术; osmFISH (cyclic single-molecule fluorescence in situ hybridization): 单分子荧光循环原位杂交; REAP-seq (RNA expression and protein sequencing assay): RNA表达及蛋白质测序; scATAC-seq (single-cell assay for transposase-accessible chromatin using sequencing): 以测序技术为平台的单细胞转座酶可及染色质检测; scBS-seq (single-cell bisulfite sequencing): 单细胞重亚硫酸盐测序; scChIP-seq (single-cell chromatin immunoprecipitation followed by sequencing): 单细胞染色质免疫沉淀测序; scGESTALT (single-cell genome editing of synthetic target arrays for lineage tracing): 使用单细胞基因组编辑人工合成芯片进行细胞谱系追踪; scHi-C-seq (a single-cell Hi-C method for chromosome conformation): 单细胞染色质构象Hi-C技术; sciATAC-seq (single-cell combinatorial indexing assay for transposase-accessible chromatin using sequencing): 以测序技术为平台的单细胞组合标记转座酶可及染色质检测; sci-CAR (single-cell combinatorial indexing chromatin accessibility and mRNA sequencing): 单细胞染色质可及性组合标记及mRNA测序; sci-MET (single-cell combinatorial indexing for methylation analysis): 单细胞组合标记甲基化检测; sci-RNA-seq (single-cell combinatorial indexing RNA sequencing): 单细胞组合标记RNA测序; SCI-seq (single-cell combinatorial indexed sequencing): 单细胞组合标记测序; scM&T-seq (single-cell methylome and transcriptome sequencing): 单细胞甲基化组及转录组测序; scNOMe-seq (single-cell nucleosome occupancy and methylome sequencing): 单细胞核小体占据及甲基化组测序; scRRBS (single-cell reduced representation bisulfite sequencing): 单细胞简化基因组重亚硫酸盐测序; scTHS-seq (single-cell transposome hypersensitivity site sequencing): 单细胞转座子组高敏位点测序; seqFISH (sequential fluorescence in situ hybridization): 序列荧光原位杂交; snmC-seq (single-nucleus methylcytosine sequencing): 单细胞核甲基胞嘧啶测序; SNS (single-nucleus sequencing): 单细胞核测序; SPLiT-seq (split-pool ligation-based transcriptome sequencing): 分裂池基于配位的转录组测序技术; STARmap (spatially resolved transcript amplicon readout mapping): 转录扩增子空间定位技术。

一、单细胞多重检测

各种单细胞分子检测技术最初关注的都是如何准确地检测单个细胞在某一状态下的某一项指标，比如最开始的半定量检测结果，以及后来的高通量DNA测序结果。最近，大家开始对单细胞多项参数的同时检测产生了兴趣，希望能够对细胞有一个综合的认识和了解。这其中往往就包括单细胞RNA测序工作以及DNA测序、蛋白质丰度检测，或者表观遗传学检测等。我们可以使用4种策略来获得一个细胞的多种数据（图2）。首先，在进行测序之前，使用非破坏性的检测；然后分离不同的细胞组份，同时进行多种试验；接下来，将多种数据进行实验转换（experimental conversion），使其成为常见的分子形式（molecular format），以便能够使用常用的方法（如DNA测序）同时检测多种数据；最后，对这些不同的数据进行分析，比如了解DNA序列、RNA丰度和序列多态性等信息。

1.

在开展破坏性试验之前先收集细胞计量信息（cytometric information）

最优的细胞全方位检测方案应该在开展破坏性试验之前，先进行非破坏性的细胞计量学检测。由于在多重单细胞RNA测序试验方案里，都会使用荧光细胞分选技术来对细胞进行各种分选操作，因此，我们很自然地会在测序之前，使用其它标记分选技术对细胞进行计量学检测（图2a）。在之前开展的细胞周期试验和半定量mRNA检测里，我们会使用细胞表面标志物，根据细胞的不同功能和状态进行分选和富集。这种方法主要在免疫学研究和造血系统研究中取得了成功，因为我们对这类细胞的表面标志物有比较深入的认识和了解。Paul和

Nestorowa这两个课题组使用这种研究策略，对啮齿类动物的造血祖细胞进行了研究，结果发现了不同的转录免疫亚型细胞。Wilson等人则使用FACS技术分离得到了罕见的造血干细胞（haematopoietic stem cells, HSC），并对其进行了单细胞RNA测序和功能检测，结果发现了与细胞持续自我更新能力有关的表面标志物。皮升（picolitre）级的新技术则有助于我们进一步提高这种试验方案的检测能力。不过，由于荧光报告因子数量上的限制，这些细胞计量学方法在单细胞水平上能够检测的参数也是有限的。

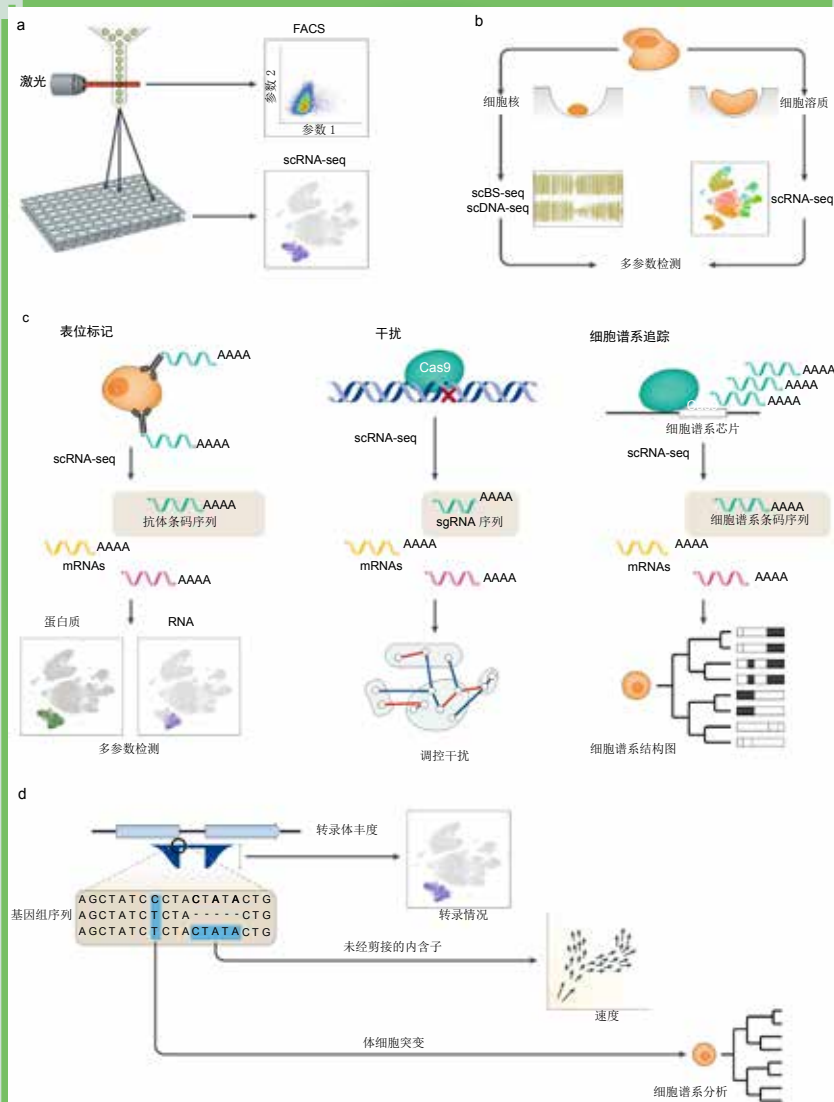


图2 单细胞多参数检测方法。a，在进行单细胞RNA测序之前，利用多通道FACS对细胞进行分选，并收集单细胞的各种参数，以利于后续使用各种荧光技术进行蛋白检测，开展转录组学研究。这样一来，我们就可以同时对单细胞进行RNA和蛋白质分析。b，使用裂解且分离的方法（lyse-and-split strategy），可以同时在不同的细胞组份进行研究。比如，将胞质与胞核分离就可以同时对mRNA和基因组DNA进行检测。c，条码标记技术与scRNA-seq技术结合，能够发现更多重要的信息，帮助我们更全面地认识细胞转录组。比如，使用多腺苷标记的抗体条码（polyadenylated antibody barcodes）来标记特异性针对细胞表面蛋白的抗体，就可以了解这些细胞表面蛋白的丰度信息（左图）。这些抗体条码序列能够与mRNA分子一同被捕获，并进入后续分析，从而帮助我们了解每一个细胞里的蛋白质丰度信息。在细胞群遗传筛查（pooled genetic screens）工作中，使用在Cas9基因编辑工作里用到的sgRNA，也可以对等位基因信息进行编码，将基因敲除信息与单细胞转录情况关联起来（中图）。在细胞谱系追踪工作中通过对谱系阵列序列（lineage array sequence）进行多次Cas9编辑，也会用到这些条码（右图）。Cas9酶多次切断谱系阵列序列，就会在其中多个不同位点形成突变。拥有相同突变的细胞最有可能源自同一个祖细胞。将这段谱系阵列序列置于RNA聚合酶II启动子的调控下，也就可以得到相应的转录产物。d，我们除了可以从scRNA-seq数据里获得每一个细胞内转录体的数量信息外，还可以发掘出其它信息，比如发现单个细胞内的体细胞突变，以及不同细胞之间的谱系关系。内含子数据则可以帮助我们估算出转录丰度的变化速度，即RNA速度（RNA velocity）。

2.

分离细胞组份

有一些细胞参数是无法通过荧光检测方法得到的，所以我们还需要其它一些检测手段。这个问题在需要同时检测一个细胞内的mRNA和基因组DNA，以及胞内蛋白质时显得尤为突出，而此时则需要对胞内的组份进行有效的物理分离（图2b）。已经有好几个实验室开发出了单细胞基因组及转录组平行测序方案，有的方案是利用生物素标记的寡核苷酸引物分离mRNA和基因组DNA，有的方案是在cDNA中选择性地插入T7启动子序列，然后在体外转录过程中选择性地扩增相应的cDNA拷贝。这就让我们对单细胞内的基因型及基因表达有了一个直观的、有关联的认识，结果发现，DNA拷贝数变异及染色体重排这两种遗传突变，可能就是导致不同细胞内mRNA丰度存在差异的原因。这些方法也对我们研究组织层面的体细胞突变非常有价值，比如用于肿瘤研究。

以Macaulay等人建立的一系列技术以及Smallwood等人开发的单细胞重亚硫酸盐测序技术为基础，并在对基因组DNA进行PCR扩增之前先用亚硫酸氢钠（sodium bisulfite）对其进行处理，这样就可以同时获得单细胞基因表达数据和DNA甲基化数据。这就是DNA多参数研究方法。由于DNA甲基化修饰是一种动态修饰，在不同的细胞之间往往都存在很大的差异，所以在分析基因表达与甲基化修饰的关系时，有必要先弄清不同细胞之间的表观遗传学差异。收集同一个细胞的DNA甲基化数据和基因表达数据，就能够了解表观遗传学与转录之间的直接关联。这也可以帮助我们了解同一细胞调控区域里的表观遗传学修饰与基因表达之间的关系，还可以更进一步证实启动子甲基化修饰会负向调控基因表达，而远离调控区域的DNA如果发生甲基化修饰，则会对基因

表达产生不同影响这一观点。使用特异性条码序列对基因组DNA和cDNA分子进行选择标记，也可以同时收集染色质可及性信息和基因表达数据。利用组合标记方法（combinatorial indexing strategy），可以同时为数千个细胞进行分析，发现这些细胞中可能对邻近基因的表达有调控作用的顺式调控原件。

对细胞裂解物进行分离的同时还可以研究同一个细胞的胞内蛋白和RNA。有研究人员利用裂解后分离方法对细胞裂解物进行分离，然后使用这些分离产物并同时对单细胞的RNA和蛋白质进行了研究。比如利用其中一种分离产物，通过邻位延伸分析技术（proximity extension assay）和定量PCR技术（quantitative PCR, qPCR）对蛋白质进行了定量研究；利用另外一种分离产物，通过定量逆转录PCR技术（quantitative reverse transcription PCR, qRT-PCR）对mRNA进行了研究。PEA技术使用两种不同的抗体来识别同一种蛋白，通过这种方法将两段成对的DNA序列拉到一起，产生了满足检测需要的强大信号。还有人使用类似的方法，将其与微流体技术结合，在同一个孔中进行逆转录和PEA反应，这就不需要对细胞裂解物进行分离了。所有这些研究也都证实了蛋白质与RNA数据的互补特点，当同时使用蛋白质数据和RNA数据时，我们可以对细胞进行更加精准的分类。不过，这些数据的数量和质量也都受到了现有技术手段的限制。目前在同一个试验中，只能同时对82~96种RNA和38~75种蛋白质进行检测（表1），因此这些技术还有很大的提升空间。此外，需要预先进行细胞组份分离，并使用微流体通路，这两者也都限制了细胞分析技术的检测通量。

3. 将细胞分析数据提炼成分子模型

将多种试验数据转换成一种单一的分子形式 (molecular format) 的试验方法是构建多元模型 (multimodal profiling) 的重要手段 (图2c), 它可以同时对多种类型的数据进行检测。比如同时将细胞表面蛋白和mRNA信息都转换成cDNA, 然后就可以利用DNA测序技术同时检测这两种不同类型的数据了。最近, 有人将DNA条码技术 (这些DNA条码上结合有特异性抗体) 与CITE-seq技术和REAP-seq技术相结合, 实现了对单细胞表面蛋白丰度和mRNA的同时检测。由于抗体条码上添加了一段poly(A)序列, 因此能够与逆转录引物分子互补结合, 并进行逆转录反应, 这样就能够捕获mRNA分子 (图2c)。这些使用DNA条码标记抗体的方法突破了使用FACS技术研究细胞表面蛋白在某些方面的局限性。使用DNA条码技术可以标记大量抗体, 理论数值可达 4^N 个 (N 表示条码序列的长度)。这些抗体能够被同时加入一个反应里, 然后通过DNA测序一一鉴别出来。而且, 这种方法同样适用于高通量的、液滴式scRNA-seq技术平台, 因此能够处理数百万个细胞, 突破了标记分选单细胞RNA测序技术的限制。与单因素分析相比, 这些多元研究也让我们能够对单个细胞有更细致的分析和了解, 比如可以对不同的免疫细胞进行更准确的区分, 这在以前, 仅仅依靠mRNA数据是不可能做到的; 还可以应用于转录后基因调控等研究。这种技术也可以用来同时研究胞内蛋白和mRNA, 不过目前还存在一定的困难, 这是因为若让条码序列进入细胞, 则需要增加细胞膜的通透性, 而这会使得胞内的mRNA分子大量降解。不过这也不是无法解决的难题, 比如可以使用PEA来检测胞内蛋白水平, 也可

以使用最近刚出现的针对固定组织的scRNA-seq技术等。

飞速发展的技术表明, 只要各种细胞内的信息能够转换成可测序的条码序列, 那么就能够在单个细胞的水平对其进行检测, 而且同时还可以对相应的转录组进行检测。这种研究策略不仅可用于研究自然状态下的细胞, 同时也可以研究处于不同非自然状态下的细胞。高通量的scRNA-seq技术和可控的CRISPR-Cas9 DNA编辑技术是非常理想的搭配, 可以对细胞进行大范围的人工干预, 进行正向遗传学研究 (forward genetic studies)。由于CRISPR-Cas9系统里使用的sgRNA分子是由RNA聚合酶III转录生成的, 所以它并未多腺苷化, 因此无法通过oligo(dT)引物进行逆转录, 也就不适用于常规的scRNA-seq平台。而CITE-seq技术和REAPseq技术则使用了多腺苷化的抗体条码, 所以可以将细胞表面蛋白信息转换成核酸序列。使用CRISPR进行细胞群遗传筛查这一技术, 使用到了多腺苷化的sgRNA条码系统, 因此可适用于标准的scRNA-seq平台。在进行单细胞mRNA测序时, 这些向导条码序列也同时被捕获, 这就可以同时了解细胞内Cas9结合位点的信息和基因表达的信息。已经有多个研究小组使用大规模的遗传干扰试验, 来研究基因调控网络 (gene regulatory networks)、未折叠蛋白反应 (unfolded protein response)、免疫细胞发育 (immune cell development) 和T细胞受体活化 (T cell receptor activation) 等问题。这种方法在研究基因调控网络方面具有极大的潜力, 虽然目前的研究工作主要集中在编码序列上, 但Cas9系统同样可以作用于基因组里的增强子、绝缘

子, 以及其它非编码序列。此外, 可经催化失活的Cas9系统还能够与不同的效应因子(比如组蛋白修饰因子、转录活化因子、转录抑制因子和DNA甲基转移酶等)结构域结合, 来研究表观遗传学修饰对基因表达的影响, 甚至可以研究敲除了多种表观遗传学效应因子之后的作用。

还有一种方法也可以将细胞信息转换成可测序形式, 那就是使用CRISPR-Cas9基因组编辑技术, 这也是最近出现的细胞谱系追踪新技术, 她能够获得许多有价值的多元单细胞数据。最近出现的多种单细胞谱系追踪技术都能够同时检测细胞内源性mRNA分子。这些方法通常都会使用CRISPR-Cas9基因组编辑技术在细胞内对细胞谱系条码序列(lineage barcode sequence)连续进行基因编辑, 然后让这些条码转录出mRNA分子, 最后使用常规的scRNA-seq对这些mRNA进行检测。每一个细胞内的条码DNA序列各有不同, 通过这些条码就可以生成一个组织或者器官进化树。考虑到不同技术在试验细节上的差异, Raj等人在前人的基础上, 使用Cas9靶标人工谱系条码追踪阵列(synthetic array), 对斑马鱼的基因组进行了编辑操作。他们将这段条码阵列置于一个可诱导的启动子控制下, 然后在收集细胞前诱导细胞转录, 并对转录产物进行逆转录操作, 最后同时检测了细胞mRNA和条码RNA。这样就能获得一份分辨率达到单细胞水平的斑马鱼大脑进化树, 同时还伴有相应的细胞转录组信息。

Spanjaard等人使用类似的方法, 对斑马鱼单细胞胚胎进行了研究, 他们在单细胞胚胎里注入了靶向16~32种红色荧光蛋白转基因(这些转基因被整合进了斑马鱼基因组的不同位点里)的gRNA和Cas9蛋白。经过一段时间, 这些红色荧光蛋白基因在Cas9蛋白的作

用下, 产生了各种突变, 形成了一系列发育信号, 最后对这些红色荧光蛋白基因的转录产物进行scRNA-seq, 如此便能了解胚胎的发育过程。

另外一种ScarTrace技术则使用了一系列H2A-GFP转基因芯片和CRISPR-Cas9编辑技术来对斑马鱼进行谱系追踪研究。在这些研究中, 科研人员往尚处发育初期的斑马鱼早期胚胎里注入了Cas9系统, 并进行了基因组编辑。待斑马鱼发育成熟之后, 使用巢式PCR技术扩增了细胞克隆里的GFP序列, 然后进行scRNA-seq。这种方法可以让我们了解细胞分化发育的多种可能性。比如有一项研究就是观察斑马鱼鱼鳍损伤之后的再生过程, 结果发现成骨细胞祖细胞克隆(osteoblast progenitor clone)会改变分化方向, 形成间充质细胞(mesenchymal cell), 再发育成鱼鳍。这只有同时掌握了单细胞的细胞谱系数据和转录组信息, 我们才能有上述发现, 这也凸显出这种技术的重要价值。

Frieda等人使用的是另外一种稍有不同的谱系追踪方法, 他们采用的是单分子荧光原位杂交谱系条码(single-molecule fluorescence in situ hybridization on lineage barcode), 而不是scRNA-seq。这种技术的优势在于能够了解处于不同空间位置的细胞之间的关系。他们使用这种光学原位读取人工突变存储(memory by engineered mutagenesis with optical in situ readout, MEMOIR)技术, 在延时显微镜(time-lapse microscopy)下对体外培养的小鼠胚胎干细胞进行了研究, 证实了这些细胞之间的关系。如果将来MEMOIR技术有更进一步的发展, 可以同时了解各种内源mRNA分子的信息, 那么它将成为一种新的多元研究手段, 帮助我们同时了解细胞谱系、空间位置和转录状态等多方面的信息。

4.

透过scRNA-seq数据发掘更多信息

虽然大部分scRNA-seq研究都在关注转录体的丰度问题，但这些研究还是可以为我们的提供许多其它方面的信息的，比如核酸序列等。这其中可能就涉及了体细胞突变、遗传变异、RNA剪切亚型等信息（图2d）。体细胞突变有可能出现在细胞基因组里的任何位置，而且可遗传给所有子代细胞，因此我们可以借此了解细胞谱系关系。已经有人通过对单个人脑细胞进行全基因组测序，利用其中的体细胞突变信息来构建神经细胞谱系关系图。更重要的是，这些人脑里的体细胞突变都出现在转录活跃的热点位置，主要都集中在编码外显子区域里，这说明神经体细胞突变是可以被scRNA-seq所发现的，也可以根据细胞转录状态来构建细胞谱系关系。此外，很多肿瘤也都会快速出现体细胞突变，这往往出现在基因组里的后期复制区域（late-replicating region）。Tirosh等人利用scRNA-seq技术对黑色素瘤细胞进行了研究，了解了其中的基因拷贝数变异情况。他们的具体方法就是对基因组里100个基因的表达值（expression values）取平均数，结果发现这些肿瘤细胞中常见的非整倍性（aneuploidy）问题。这就让单细胞基因型信息与细胞转录组信息结合起来，可以更全面地认识肿瘤细胞。Fan等人也利用scRNA-seq技术，结果发现了细胞基因拷贝数变异和杂合性丢失（loss of heterozygosity, LOH）现象；并且用这种方法对多个黑色素瘤患者的肿瘤组织标本进行了研究，发现了不同肿瘤克隆之间存在转录异质性（transcriptional heterogeneity）现象。

单细胞分析还可以让我们用一种全新的方法来认识DNA自然变异对基因表达和细胞状态等细胞表型的影响。Kang等人开发了一

套全新的scRNA-seq数据分析工具，对不同来源的细胞进行基因分型分析。他们对23个不同供体来源的人体细胞进行了全基因组关联研究（genome-wide association study），结果发现了不同细胞之间与细胞类型特异性基因表达差异有关的表达定量性状位点（expression quantitative trait loci, eQTL）。他们还发现了与免疫细胞形成不同构成比例有关的遗传变异，这也说明这种方法能够很好地将基因表达和细胞类型频率（cell-type frequency）作为定量数据，并与遗传变异联系起来。van der Wijst等人也使用了类似的方法，并且将来还会开展更大规模的研究，这将进一步提升这类eQTL研究的分辨率，更加精确地发现与细胞状态和功能有关的遗传变异。

更重要的是，我们还可以根据scRNA-seq结果（哪怕是那些只能捕获RNA 3'端的测序技术）了解与转录亚型有关的关键信息，以及内含子保留信息。这些信息的丰富程度足以让人大吃一惊，对转录丰度研究具有极大的补充作用。La Manno等人开展了一项开创性的研究工作，他们发现未剪切内含子的频率与mRNA生成降解比有关，在越“新鲜的”转录产物中，含有内含子的可能性越高。因为这些RNA在充分剪切成成熟的mRNA之前，被捕获的可能性也越高。La Manno等人通过检测未剪切内含子的频率，就能估计出转录丰度的变化速率，即RNA速度（RNA velocity），以及每一个细胞将来的转录状态。我们还可以利用拟时轨迹技术（pseudotemporal trajectory）预测出细胞未来可能呈现的状态，解决使用其它单细胞拟时检测技术无法解决的难题，比如确定一段时间的起始点和细胞周期等。这些方法都有望彻底改变单细胞生物学的现状。

5.

分析多元数据

快速发展的多元分析策略让我们意识到目前迫切需要一大批创新型工具，来对这些数据进行多元分析。虽然这些技术目前大部分还都在研发之中，但我们相信多元数据可以帮助我们发现细胞状态之间更加细微的差别，这是依靠单一的单细胞数据是无法发现的（图3a）。scRNA-seq数据的情况尤为突出，因为scRNA-seq技术不能充分认识低表达基因，所以就无法精确地进行判断，但是利用其它单细胞数据便可以弥补此不足，并能解决这个问题。比如，仅利用scRNA-seq就很难区分记忆T细胞和调控T细胞等亚型T细胞，但是再结合细胞表面蛋白标志物，就很容易鉴别了。这说明联合同一细胞的免疫表型（immunophenotype）和mRNA水平数据，可以非常清晰地鉴定免疫细胞的状态（图3b）。联合同一细胞的转录组、染色质组或甲基化状态等数据，我们能了解不同细胞在基因表达调控方面的差异，这些调控机制能够在转录前水平决定细胞的最终命运。

在单细胞多元数据分析中我们还可以联合使用统计工具，已有研究者使用这种策略对多种组学数据进行了降维处理，并发现了其中的共同点或差异点，这种方法也可以应用于单细胞多元数据分析工作。比如，Argelaguet等人开发的多组学因素分析方法（multi-omics factor analysis, MOFA）就能够发现哪些因素可以决定不同数据之间的差异。他们使用这套技术对多位慢性淋巴细胞白血病（chronic lymphocytic leukaemia）患者的基因组学、DNA甲基化和RNA表达数据进行了联合分析，结果发现转录活化决定了每位患者的体细胞突变状态。研究人员使用MOFA方法对87个小鼠胚胎干细胞的单细胞甲基化数据和转录组数据进行联合分析，结果发现这些数据之间具有明

确的关联。这些干细胞在向定向祖细胞分化的过程中，DNA甲基化水平和转录组的变化之间都有明确的关联。这些发现表明，MOFA等方法非常适合分析多元单细胞数据，而且有助于我们将来对这些数据进行整体解读。随着单细胞数据的数量远远超过样品的规模，我们更加需要强有力的、能从多个角度开展机器学习（multiview machine learning）的手段，只有这样，才有可能开发出有价值的单细胞分析技术。

开展多元单细胞试验，也是一个了解细胞内不同数据之间关系的好方法。比如，同时检测同一细胞的mRNA和蛋白质数据，就很容易发现细胞内相关性比较差的时刻，即活性转录后调控阶段。再进一步结合其它因素，比如RNA速度，就能够绘制出内容丰富的蛋白质和RNA生成动态图。此外，我们还可以利用统计模型来评价研究结果的准确性，比如在模型中修改某一参数，看看是否会对其它参数带来预计的改变等（图3c）。Cao等人就利用线性回归模型（linear regression model），根据染色质可及性数据预测了基因表达值。结果发现使用这套系统进行预测的准确性要比仅使用顺势调控元件数据进行预测时高得多，比如在预测远隔位点处的染色质可及性的作用时，准确性提高了四倍。Dixit等人构建的正则化线性模型（regularized linear models）能够预测哪些gRNA对每一个细胞的基因表达水平有何影响，这样就可以明确细胞反应的决定因素，重构转录网络。这些模型可以用来判断细胞内哪些基因表达的改变是gRNA，或者其它因素导致的。因此，我们预测多元数据将能帮助我们更加明确细胞内复杂调控网络的作用机制，包括表观遗传调控、转录调控和转录后调控等。

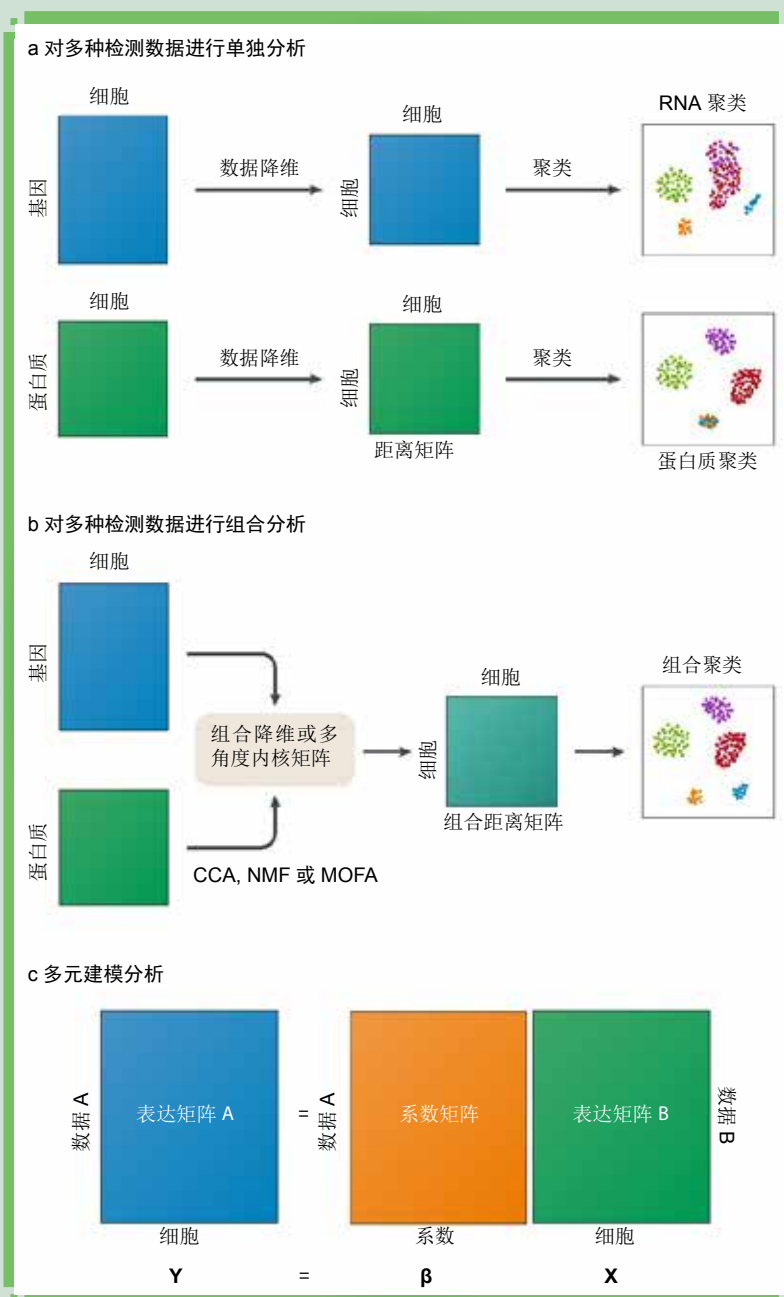


图3 单细胞多元数据计算方法。**a**，对同一群细胞的不同检测数据进行单独分析，这可能会得出完全不同的结果。比如使用基因数据进行分析时，可能会认为这一群细胞大致是相似的，但使用蛋白质数据进行分析时，则完全不同。图中橙色和蓝色的细胞在以转录数据进行分析时可以看出明显的差异，但是在以蛋白质数据进行分析时，就无法分辨了。红色和紫色的细胞在RNA层面无法区分，但是在蛋白质层面就有明显的区别。**b**，对同一群细胞的不同检测数据进行联合分析，就能够更准确地区分出不同的细胞聚类。这比以单一数据为分类标准更加准确。如图中所示，联合蛋白质和RNA数据，就可以将红色、紫色、橙色和蓝色细胞进行很好的区分。**c**，利用统计方法，可以了解不同检测数据之间的关系，比如了解两种数据之间的协变量关系。CCA（canonical correlation analysis）：经典相关分析；MOFA（multi-omics factor analysis）：多组学因素分析；NMF（non-negative matrix factorization）：非负矩阵分解。

二、整合不同单细胞试验获得的数据

前面，我们重点介绍了同一个细胞不同数据的整合方法，可是在单细胞生物学领域，如何将不同单细胞的数据整合起来，才是一个大难题。另外，还需要解决大量数据集之间的“批次校正（batch-correction）”问题，这也是妥善利用不同实验室和不同试验流程得出结果的先决条件。可是现有的手段都无法适用于单细胞数据分析，因为无法区分不同种类细胞在组份比例上的区别，以及细胞内分子过程上的变化。所以我们迫切需要开发出新的单细胞批次校正技术，而这首要的便是需要能够识别出不同数据集之间共有的生物学状态，比如匹配的细胞种类（图4a）。目前，这个研究领域也已经成为了一个飞速发展的领域。

1.

整合scRNA-seq数据

我们最近介绍的一种Seurat v2 R软件就可以对不同数据集里的单细胞数据进行校正。这套软件首先使用经典的相关性分析方法，找出不同数据集之间共有的变异来源（图4a）。由此得出的经典相关向量就代表了不同数据集间的共有细胞种类，这就可以代表批次效应（batch effect），或者数据集特异性的异质性来源，而这些因素都可以通过标准的主成分分析方法来捕捉（图4b）。接下来，借助动态

时间校正这种非线性转化技术，将这些源自不同数据集的经典的相关向量进行对齐，我们就能够修正不同细胞群密度数据之间的差异了。经过这两步处理，细胞就进入了一个低纬度空间，而这个低纬度空间是不同scRNA-seq数据集共有的。在这个空间里，不论采用了哪种试验方法，进行了哪种试验，那些拥有相同生物学状态的细胞都聚集在一起了。

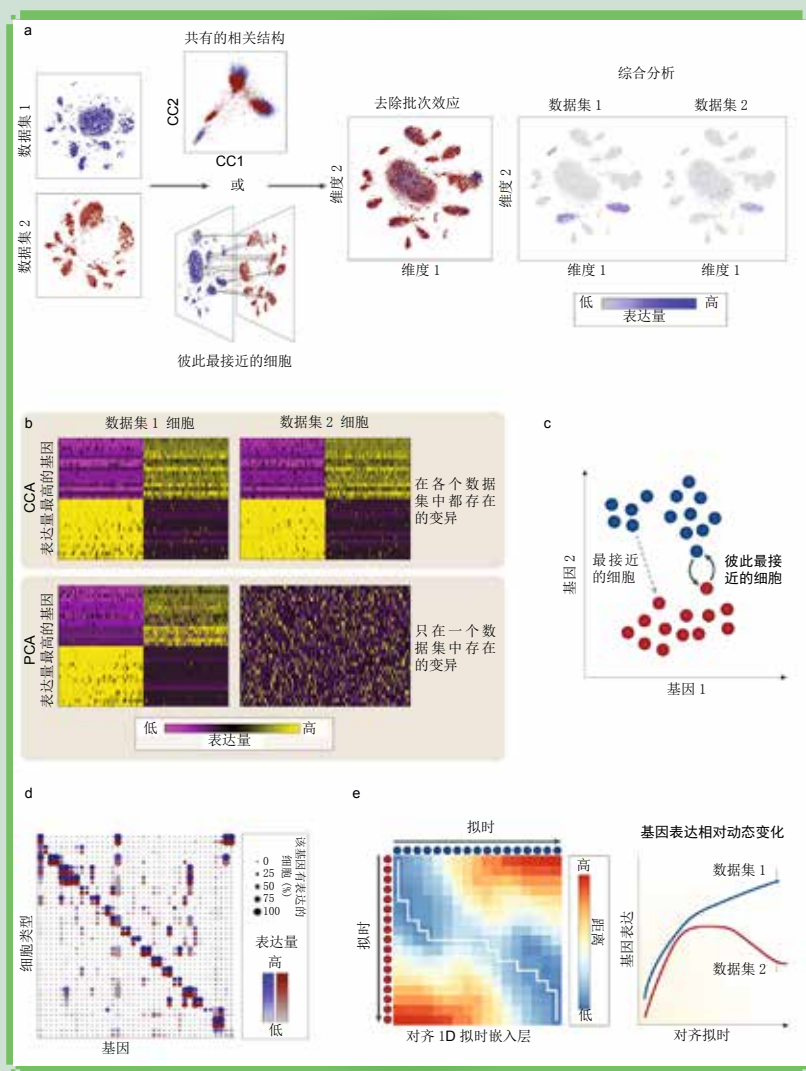


图4 单细胞多元数据整合分析软件工具。**a**，利用计算机软件，能够将多种不同的数据集整合在一起进行综合分析，这也有利于后续的比较分析。通过经典的相关性分析或者彼此最相邻分析，就可以发现共同的相关结构。在不同的细胞之间找出共有的空间（shared space），或者等效细胞（equivalent cells），就可以消除批次间的差异，从而在不同的细胞之间直接进行比较。**b**，经典的相关性分析可用来发现在两个数据集之间最大的相关变量。相比之下，主成分分析等方法则可以发现决定单一数据集里最大变量的两个正交变量（orthogonal variables）。**c**，通过细胞的基因表达谱来发现彼此最相邻的细胞，就可以确定在生物学上等效的细胞。找到了这些细胞之后，就可以对不同数据集的基因表达数据进行校正，计算出表达数据的转换结果。**d**，明确了等效细胞状态之后，就可以在不同的数据集之间，对每一个细胞状态的基因表达值进行比较，发现不同数据集之间基因表达的异同。此时，每一个数据集里的细胞都按照细胞类型进行了分类，而且每一种细胞的目标标志物的表达量也做了标注。图中每一个点其大小都与细胞群中表达了该基因的细胞所占的比例一一对应，每一个点的颜色则与基因的表达水平一一对应。从图中可以一目了然地看出不同数据集里的同一种细胞，彼此之间同一基因的表达差异。利用统计学方法还可以发现不同数据集内同一种细胞（或同一群细胞）基因表达上的差异。**e**，将不同数据集的1D拟时向量进行对齐操作，可以去除细胞动态过程之间在时间上的差异，确定两个数据集里相同的时间点。然后直接比较基因表达数据，就可以发现两者在基因表达方面的异同。CC（canonical correlation vector）：经典的相关向量；CCA（Canonical correlation analysis）：经典的相关性分析；PCA（principal component analysis）：主成分分析。

mnncorrect是另外一种校正方法，它主要利用了图像识别和匹配技术。这种方法运用的策略是找出彼此最接近的细胞（**mutual nearest neighbours, MNN**）（图4c），这些细胞也最有可能具有共同的生物学状态。这些匹配的**MNN**细胞之间的距离可以用来计算批次向量，然后校正不同数据集里的基因表达数据。更重要的是，使用**MNN**方法无需每一个细胞都有一个**MNN**，因为在这种方法里，我们可以通过计算相邻匹配细胞的批次向量，来得出每一个细胞的批次向量。

不论是使用**CCA**方法还是**mnncorrect**方法，都可以对不同实验室，及不同试验方法得到的**scRNA-seq**数据集进行整合分析，只要这些数据都来自同样的组织就行了。有研究者使用这两种方法对文献里的数据集（4个人体胰岛单细胞数据集）进行了再分析，结果成功获得了一个整合数据集，并用这个更大规模的数据集进行了荟萃分析。这种统计效力更高的分析帮助我们发现了更加罕见的细胞状态，以及与之对应的基因表达标志物和转录亚型。因此，这类技术也有望成为第一个帮助我们解决如何校正不同实验室及大型科研项目数据的关键手段，进而可利用众多单细胞试验数据构建出一个单一的参照数据集。由于这些工作具有极大的应用潜力，所以开发**scRNA-seq**结果整合技术已经成为了一个新兴的、发展迅速的方向。目前，大家的工作重点主要都集中在提高软件的效率方面。尤其值得一提的就是**Korsunsky**等人开发的新型**k**均值聚类方法（**k-means clustering**），这种方法能够将不同数据集里的细胞进行聚类，可以利用个人电脑对50万个细胞进行整合分析。

更重要的是，对这些单细胞数据集进行有效整合，不仅能解决批次校正的问题，还能对不同生物条件下的单细胞进行深度比较（图4d）。比如，**Butler**等人对不同试验中、受 β 干扰素刺激的实验组和没受 β 干扰素刺激的

对照组人体外周血单核细胞数据集进行了整合分析，结果发现了13种共有的细胞类型；同时他们也对这些细胞的转录组进行了全面比对，发现了与 β 干扰素刺激有关的细胞类型特异性反应（**cell type-specific responses**）。这一研究工作发现，浆细胞样树突状细胞（**plasmacytoid dendritic cell**）对 β 干扰素刺激有一种特有的强烈反应，而借助流式细胞技术和大规模**RNA**测序也能检测到相应的、特异性的基因单元（**gene module**）。我们估计，类似的分析工作有望发现与环境 and 遗传干扰有关的细胞类型特异性反应，甚至能够对不同疾病和不同治疗方案下的患者样品数据进行标准化的比对。

对大量**scRNA-seq**数据进行整合不仅会涵盖各种科研技术和条件下的试验数据，还会涉及到跨物种的分析。实际上，多个研究小组早就已经使用这种方案来研究不同系统里的细胞状态进化问题。比如，**Karaikos**等人就对种类不同的两种果蝇在胚胎发育早期阶段的基因空间表达模式（**spatial gene expression pattern**）进行了比较。他们的研究方法是绘制空间基因表达图谱，借此就可以对不同物种垂直同源基因（**orthologous gene**）的表达情况进行全面比较，查看其在进化上的变化情况。**Tosches**等人则对爬行动物的脑细胞进行了**scRNA-seq**研究，并计算了两者平均基因表达值之间的相关性，借此与哺乳动物的脑细胞结果进行了比对，结果发现了在这两者间保守的抑制神经细胞，以及不保守的激动神经细胞。**Baron**等人构建了小鼠和人的胰岛**scRNA-seq**数据集，结果发现了两者之间匹配的细胞和调控机制，尤其值得一提的是发现了四种高度保守的内分泌细胞。这些保守的细胞也可以在无监督的情况下，用**Seurat CCA**比对方法进行分析。

Alpert等人开发的**cellAlign**方法能够对人和小鼠的胚胎发育数据集进行比对（图4e），

他们的策略就是在1D维度上进行拟时轨迹对齐，结果发现人体胚胎在合子（zygotic）阶段的基因组活化要晚于小鼠。在小鼠细胞内，与蛋白质合成有关的基因的“动作”都要快一些。虽然这些方法目前还都处于初级阶段，但都已经表现出了极大的潜力，将来都可以在单

细胞水平用于比较生物研究工作，发现模式生物细胞与人类细胞的相关性。在经历了数百万年进化的不同物种间，准确发现垂直同源基因是进行可靠数据整合的前提条件，也将是未来面临的巨大挑战。

2. 将不同scRNA-seq数据集里的细胞进行分类

一直以来，科研人员都希望能够将试验里的细胞类型分类（cell type classification）信息在不同的试验中进行传递，因为这要比从头进行细胞聚类操作更有效（图5a）。随着高质量的细胞注释信息变得越来越多，这种需要也将变得更加强烈。迄今为止，有两种方法能够将细胞一一对应到已有的数据集里，实现细胞标记信息的传递，那就是scmap-cell方法和scmap-cluster方法，这两者之间也互有关联。第一种方法能够发现不同数据集里最邻近的细胞，可以根据邻近细胞上的标签，将细胞按照类型进行分配，打上细胞类型标签（cell-type label）。第二种方法是通过查询参考数据集这一方法，在其中找到最接近的细胞聚类，来对细胞进行分类。这个参考数据集是根据相关性距离

（correlation-based distance），或者余弦相似度（cosine similarity）来组建的。正在开发中的新技术则使用了奇异值分解（singular value decomposition）、线性判别分析（linear discriminant analysis），或支持向量机（support vector machines）等技术，以注释参考数据为依据，对细胞进行分类。需要提醒的是，与无监督聚类操作（unsupervised clustering）相比，监督注释（supervised annotation）下的分类要更加准确，尤其在参考数据的样本量、测序深度、覆盖率不断增加的情况下更是如此。利用参考数据指引下的细胞注释来明确细胞亚群，有助于对相关细胞进行分析，这是通过从头聚类方法无法实现的（图5b）。

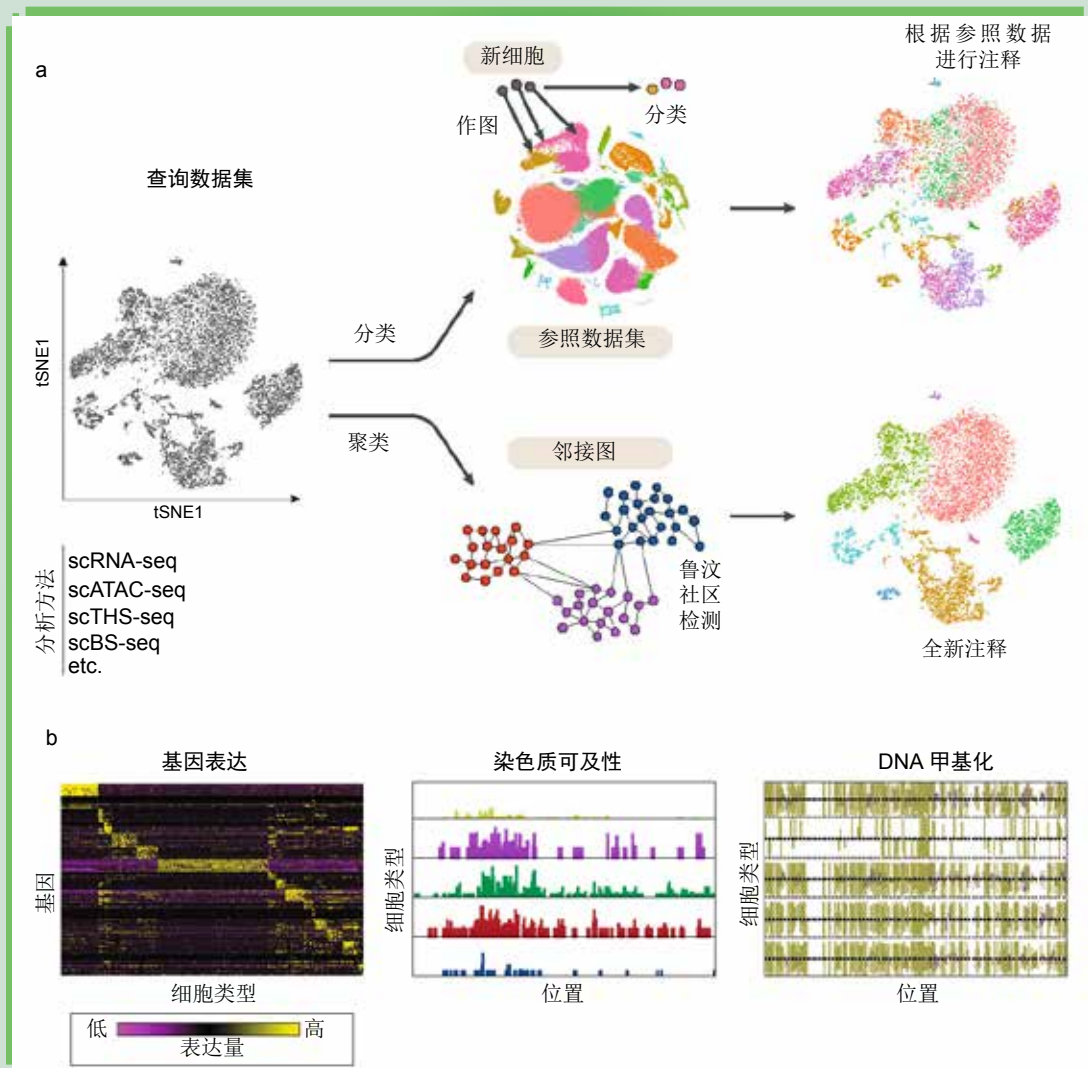


图5 细胞聚类及分组分析。**a**，有两种方法可以对一个全新数据集里的细胞进行分类。首先，能够以之前类似细胞的数据做参照，来进行分类，也可以用无监督聚类法（**unsupervised clustering**）来发现相似的细胞。根据参照注释进行分类时，如果借助更高质量数据库里呈现出的结构，结果就会更加准确。**b**，不论是借助参照数据，还是从头开始进行全新分类，只要是根据常规的特征对细胞进行分类，细胞之间的差异还可以通过不同的方法进行后续分析。比如利用基因表达数据来发现在不同类型的细胞，或者不同群体的细胞之间表达有差异的基因。利用染色质可及性数据来发现基因组里可及性不同的区域，以及富含DNA基序（**DNA motifs**）的位置。利用单细胞DNA甲基化数据来发现不同类型的细胞，或者不同群体的细胞之间，甲基化修饰有差异的位置等。tSNE（**t-distributed stochastic neighbour embedding**）：t-分布随机邻域嵌入；scATAC-seq（**single-cell assay for transposase-accessible chromatin using sequencing**）：以测序技术为平台的单细胞转座酶可及染色质检测；scTHS-seq（**single-cell transposome hypersensitivity site sequencing**）：单细胞转座体高敏位点测序；single-cell bisulfite sequencing（**scBS-seq**）：单细胞亚硫酸氢盐测序。

3.

多元单细胞数据的整合

将scRNA-seq数据与其它没有共同特征的单细胞数据进行整合，是另外一个完全不同的问题，需要其它手段和方法来解决。比如比较染色质可及性或DNA甲基化修饰等基因组数据和基因或蛋白质表达数据，因为这两者之间的相关性就不是那么清晰。不过，在不同类型的数据里，从相似的细胞群里收集细胞，还是可以找到共同的生物学状态，通过这种方法也可以发现多元数据之间的相关性。Welch等人就开发了一种MATCHER方法，能够对不同试验里的1D拟时轨迹（1D pseudotime trajectories）进行比对，而且还能够对不同类型的数据进行比对。MATCHER方法假定有一条共同的发育轨迹，对不同的数据都有同样的影响，然后让不同试验里的细胞全都对应到一个共同的1D拟时空间（common 1D pseudotime space）上。这种对应投射能够发现不同试验里的细胞，而不需要提前知道这些试验结果里的相关特征。Welch等人使用MATCHER方法将人诱导多潜能干细胞（induced pluripotent stem cell, iPSC）的单细胞RNA测序数据与单细胞甲基化组及转录组数据进行了整合，研究了这些细胞在重编程过程（reprogramming）中，转录组和DNA甲基化修饰的动态变化情况。结果发现DNA甲基化修饰常常是滞后于基因表达改变的。

虽然我们常常借助scRNA-seq数据，通过其中的细胞特异性标志物基因（cell-type-specific marker gene）的表达情况来确定细胞的种类，但是对于染色质可及性区域（accessible chromatin region）等信息与细胞种类之间的关系还不是非常清楚。将这类信息与scRNA-seq结果和细胞分类结果整合，就能够帮助我们了解这些数据与细胞种类

之间的对应关系了。Lake等人就将单细胞核RNA测序（single-nucleus RNA sequencing, snRNA-seq）结果与单细胞转录组高敏位点测序（single-cell transposome hypersensitivity site sequencing, scTHS-seq）结果相结合，对一系列相匹配的大脑组织切片进行了研究。他们以单细胞基因表达数据为指引，使用梯度提升技术，将不同种类的细胞与各自的染色质可及性数据一一对应起来。他们首先分别在scRNA-seq和scTHS-seq的结果里确定了相应的细胞类型；然后训练机器模型，使其将基因表达情况与染色质可及性对应起来；之后再使用这套模型对其它scTHS-seq细胞进行分类。这种将多元数据进行整合的分类方法能够对大脑染色质可及性数据进行更加精细的解读，包括发现与很多常见遗传疾病有关的病理细胞类型等。使用同样的方法，还可以对单细胞DNA甲基化数据或单细胞转座子可及性染色质测序（single-cell assay for transposase-accessible chromatin using sequencing, scATAC-seq）数据进行整合分析。

正在开发的其它新方法则可以对细胞进行多元数据交叉分类（cross-modality classification），这些方法假定细胞都有一个共同的特征，或者是一个在不同数据之间共有的特征。Welch等人开发的LIGER技术就是一种整合式非负矩阵分解技术（integrative non-negative matrix factorization, iNMF），可用于这类分析工作。他们使用LIGER技术，利用单细胞亚硫酸氢盐RNA测序数据对大脑皮质细胞进行了分类。他们假定基因体甲基化与基因表达之间存在负相关关系，然后整合了多元数据，最后对细胞进行了联合聚类（jointly

clustered) 操作。我们最近还介绍过Seurat v3整合分析软件, 该软件能够对拥有不同检测数据的细胞进行分类, 其策略也是假定不同检测数据间拥有相同的或相关的特征。这就能够根据来自同一组织的scRNA-seq数据注释信息, 借助scATAC-seq结果对细胞进行分类,

并发现单独依靠scATAC-seq结果无法发现的细胞亚类结果。我们希望将来借助这些技术, 对各种单细胞数据进行更细致的分析, 在更多的组织中, 发现更多与细胞类型相关的染色质可及性特征和DNA甲基化特征。



资讯 · 频道

www.LifeOmics.com

三、将测序结果与空间数据相结合

细胞在组织里的空间组织信息往往折射出不同细胞在功能上的区别，以及细胞谱系和最终命运间的差异。基因表达在空间上的差异也增加了给发育中的多细胞器官做注释的复杂程度，因为不同基因的表达决定了细胞不同的分化途径，也决定了这些细胞在空间上的精确排布，并最终决定了整个组织和器官的形成。更重要的是，我们无法通过细胞的RNA表达谱来获得这些空间信息，因为进行RNA测序之前，都会先破坏组织，使细胞溶解、释放出来。可能具有相似基因表达谱的等效细胞（Equivalent cell types）能够在原位保持特定的空间结构，因此，丢失空间结构信息成为很多单细胞分析手段的致命缺陷。将细胞的空间结构信息与基因表达信息相结合，则有望解决这个问题（图6a）。我们可以借助计算机软件，或者同时收集细胞的原位空间坐标信息和基因表达值信息，来达到这个目的。

目前，已经有好几种方法达到了原位（保留空间信息）检测基因的表达水平。比如，使用荧光原位杂交技术，该技术是原位检测基因表达的金标准。而且迭代之后更新的技术，已经使检测效率接近100%。荧光原位杂交技术常常用于在一个实验中对少量细胞进行观察，通常无法对一个细胞内的所有表达基因都进行检测。不过，最新的技术已经结合了有纠错功能的测序杂交探针，能够在一次实验中对数百个基因的表达进行检测，也可以结合空间条码技术，在进行mRNA逆转录的过程中，记录相应的空间信息。相比之下，scRNA-seq技术则无法收集空间信息，但是却可以一次收集单个细胞内的数千个转录子信息，一次实验同时处理数亿个细胞。通过软件将这些不同类型的数据整合在一起，就能够发挥每一种检测手段的优势，既能够利用scRNA-seq的高通量优势，也保留了相应的、高分辨率的位置信息。

使用荧光原位杂交技术与scRNA-seq技术相结合，已经让科研人员取得了多个标志性的成果。比如最早的Satija课题组和Achim课题组发表的文章，以及后来对其它组织进行的类似研究都在器官或组织水平，绘制出了基因表达空间结构图谱。这些研究都探索了那些已知具有空间分布模式的关键基因的空间分布情况，并用这些数据为每一个基因构建了表达模型。反过来，这些基因的空间表达模型又可以用来对已经经过scRNA-seq的单个细胞进行空间作图，具体方法就是以具有空间标志物的基因表达情况为依据，进行空间定位。在最终的基因空间表达模型里，几乎每一个基因都有一份高分辨率的空间结构信息，每一种细胞周围都有哪些细胞，也都一目了然（图6b）。目前使用这些方法已经发现了一些受空间位置调控的基因，也拓展出了新的研究领域（图6c）。未来，还将能够对基因的空间表达趋势进行系

统性的分析。但是，荧光原位杂交与**scRNA-seq**技术相结合的研究手段，目前主要还是应用于空间结构已经非常清晰的组织和器官的研究，比如早期胚胎和哺乳动物的肝脏等。如果要对组织切片、实体瘤等空间结构更加复杂的对象进行研究，还存在一定困难。还有一些研究也已经在其它方向上做了一些粗浅的尝试，他们也是将基因的空间表达数据与**scRNA-seq**结果相结合，具体的方法就是利用荧光原位杂交，或免疫组化等方法确定目标细胞群，然后对其中的少数几个目标基因进行检测。不过，对于空间结构更为复杂的组织和器官，我们还无法像对空间结构简单的组织和器官那样，进行多元数据整合。

最近，业界又出现了两种高分辨率的空间基因表达检测方法，它们都能够在更大的2D和3D结构上，对单细胞里的数千个基因进行检测。这些方法都极大地降低了检测组织的背景荧光噪声，从而提高了基因检测的信噪比。比如第一种方法使用了循环单分子荧光原位杂交技术（**cyclic smFISH**, **osmFISH**），组织切片能够与盖玻片共价结合，从而能够对组织进行更好的透明化处理。另外一种方

法则使用了空间解析转录扩增子数据作图技术（**spatially resolved transcript amplicon readout mapping**, **STARmap**），在探针的原位扩增过程中（逆转录过程），将人工修饰过的DNA碱基插入其中，使cDNA产物能够与聚丙烯酰胺基质（**polyacrylamide matrix**）共价结合，从而对组织进行更高标准的透明化处理，而又不会丢失3D结构信息。这些方法都能对基因的表达情况进行准确的原位检测，因此，可以对原位细胞进行更加精确的分子分型，同时掌握它们的空间结构信息（图6a、b）。科研人员分别利用上述两种方法，对小鼠皮质进行了研究，并根据多种细胞标志物基因的表达情况，对这些皮质细胞进行了准确的分类。更重要的是，他们也同时分析了这些细胞的空间定位信息，这样就能按照解剖学的结构，对这些细胞进行区分，并能够在3D的立体空间尺度上，分析不同类型细胞的分布情况。将空间数据与**scRNA-seq**结果，及其它单细胞检测结果进一步整合，将创造出一个前所未有的新机遇，从而让我们更加清楚地认识组织的组成和功能。

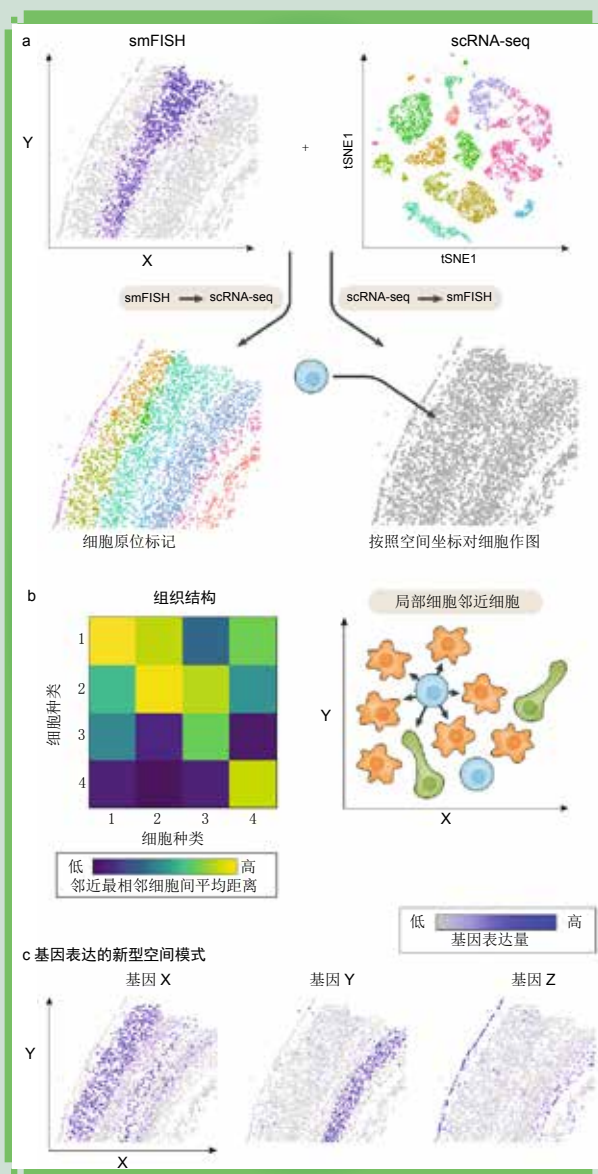


图6 单细胞空间数据整合分析。a，有两种方法可以将scRNA-seq结果与单分子荧光原位杂交（single-molecule fluorescence in situ hybridization, smFISH）结果整合起来进行综合分析，分别是使用smFISH数据对scRNA-seq结果作图，使用scRNA-seq结果对smFISH数据作图。使用第一种方法，能够使以RNA测序结果为依据的细胞分类转换成细胞在空间上的分类，如左图所示。而使用第二种方法，能够使彼此在RNA测序结果上无关联的细胞“回归”各自的空间位置，如右图所示。b，细胞完成空间整合之后，就可以进行组织结构分析，了解组织的细胞组成，以及这些细胞彼此之间的空间位置关系。有一种方法可以了解不同种类的细胞在组织里的空间组织结构，那就是查看每一种细胞的周围存在哪些细胞。通过衡量不同种类细胞之间的平均距离，便可以了解组织的结构信息，比如哪类细胞在组织里是分散存在的，哪类细胞在组织里是相邻存在的（这些彼此靠近的细胞之间往往都存在相互作用）。c，综合smFISH数据，用经过scRNA-seq的细胞对细胞空间坐标作图，可以预测细胞里每一种基因在空间上的表达情况。使用这种预测模型，我们甚至可以预测没有相应smFISH数据的基因的表达情况。smFISH（single-molecule fluorescence in situ hybridization）：单分子荧光原位杂交。tSNE（t-distributed stochastic neighbour embedding）：t-分布随机邻域嵌入。

四、前景展望

随着各种单细胞研究技术的不断发展和成熟，不论是数据的规模，还是质量，都在飞速增长。因此，我们也更加迫切地需要一种单细胞多元数据整合手段。目前，已经有大规模的协作项目在进行当中了，比如人类细胞图谱计划（**Human Cell Atlas**），就计划为人体内的每一个细胞构建一份完整的图谱，同时也有其它一些类似的项目，计划对重要的模式器官构建同样的数据库。这些数据库不论是构建，还是使用，都离不开有效的数据整合手段的支持。首先，如果要将不同实验室得到的数据、不同实验手段得到的数据和不同检测对象的数据整合到一起，就需要有效的方法来消除彼此之间的差异。其次，还要使数据能够在不同的图谱项目间，以及不断新出现的数据间进行传递和比较分析。从项目意义上说，这就有点类似于人类基因组计划，正是因为有了这个项目，才使后续的科研工作更加方便，而且也极大地降低了后续研究的成本。我们还希望，将来能够对单细胞进行更多方面的检测。纳米孔测序技术（**Nanopore sequencing**）在多元单细胞应用领域就具有极大的潜力，因为该技

术能够直接对**DNA**和**RNA**进行长距离测序，并同时分析剪辑的修饰情况。随着科技不断的发展，将来还会看到更多的新技术，比如检测蛋白质的新技术等。不断发展的细胞高分辨率空间检测技术，也将帮助我们在细胞的原位对其进行观察，掌握更多细胞在组织内排列的信息。总而言之，将不同实验条件下的、单个细胞内的多种不同数据整合起来，可以帮助我们跳出以转录组为中心的科研模式，对细胞和组织、器官有一个更全面的认识和了解。通过研究同一个细胞内，不同数据之间的关系，我们也可以揭示不同细胞功能背后的机制，了解它们彼此之间的因果关系。在生物学研究领域，有一个最主要的问题就是：“什么是细胞类型（**what is a cell type**）？”如果这个问题真的有答案，那肯定只能通过对单细胞进行细致的研究，将不同实验条件下的不同信息综合起来，才能找到答案。这就好像另外一个经典的问题：什么是基因（**what is a gene**）？要回答这个问题，也只有通过对不同物种进行全面、细致的**DNA**测序，同时结合其它生物化学检测的结果，才能给出正确的答案。

名词解释:**1. Single-cell RNA sequencing (scRNA-seq).**

Sequencing of cDNAs derived from RNA molecules (usually polyadenylated mRNAs) from a single cell. it is typically performed for many hundreds to thousands of cells in a single experiment.

单细胞RNA测序，即对单细胞内的mRNA等RNA分子的逆转录产物cDNA进行测序，一次试验里通常都会对几十万个细胞进行检测。

2. Multimodal

Data of multiple types, for example, of RNA and protein.

多元检测，即对多种数据进行检测，比如RNA和蛋白质等。

3. Index sorting

Fluorescence-activated sorting of cells into known plate locations.

标记分类，对细胞进行荧光活化分选。

4. In vitro transcription

Transcription of a DNA sequence in vitro using the T7 RNA polymerase.

体外转录，利用T7 RNA聚合酶在体外对DNA序列进行转录，生产相应的RNA。

5. CITE-seq and REAP-seq

Cellular indexing of transcriptomes and epitopes by sequencing (CiTe-seq) and RNA expression and protein sequencing assay (ReAP-seq) are methods that are capable of detecting cell surface protein abundance and gene expression within the same single cell. They achieve this through the use of barcoded antibodies captured alongside mRNA transcripts in single-cell RNA sequencing (scRNA-seq) experiments.

使用基于测序技术的细胞转录组及表位标记技术和RNA表达及蛋白质测序技术，能够检测细胞表面蛋白的丰度，以及同一细胞内基因的表达情况。其中还需要借助条码抗体的帮助，在进行单细胞RNA测序时捕获这些mRNA分子。

6. CRISPR-Cas9

A protein-RNA complex that allows targeted mutation or binding of DNA sequences as determined by a guide RNA sequence.

CRISPR-Cas9复合体是一种由蛋白质和RNA分子组成的复合物，能够在向导RNA的指引下，靶向特定的突变，并与DNA分子结合。

7. Pooled genetic screens

Screening experiments in which each individual cell may receive a different perturbation at random without prior separation of groups of cells and perturbation treatments.

细胞群遗传筛查，这是一种针对大量细胞，在不需要进行预先分选和其它处理的情况下，可以集中进行遗传筛查的技术，其中每一个细胞都会随机受到不同的处理。

8. Lineage tracing

The identification of lineage relationships between groups of cells through shared DNA mutations.

细胞谱系追踪技术，能够根据不同的细胞是否携带有相同的DNA突变，来判断这些细胞彼此之间的“亲缘关系”。

9. Single-molecule fluorescence in situ hybridization (smFISH).

A fluorescence in situ hybridization method capable of detecting the presence of a single molecule (usually RNA) through the recruitment of many fluorophores to the same area. it enables a quantitative readout of the number of molecules present in a cell.

单分子荧光原位杂交技术，这种荧光原位杂交技术可以将大量的荧光信号汇集到一处，因而能够对RNA等单个分子进行定量检测。

10. Expression quantitative trait loci(eQTIs).

genomic loci that explain variation in the RNA expression levels of genes.

表达数量性状位点，即能够解释基因RNA表达水平差异的遗传位点。

11. Intron retention

The presence of intronic RNA bases in an RNA transcript. These bases are usually removed by RNA splicing shortly after or during transcription.

内含子保留，在RNA分子里还保留有内含子的碱基，但是在转录过程或RNA剪切过程中，通常都会去除这些碱基。

12. Pseudotime

The ordering of cells along a one-dimensional axis describing a continuous differentiation process.

拟时间，延一个轴向（比如时间），将细胞按照一个连续的过程进行先后排序。

13. Joint clustering

grouping cells on the basis of measurements from multiple data modalities.

联合聚类，根据多种检测数据对细胞进行分类

14. Canonical correlation analysis (CCA).

A statistical method for investigating relationships between two data sets. CCA aims to identify shared sources of variation in a pair of data sets.

经典相关分析，这是一种统计分析方法，能够了解两组数据之间的关系。主要用来判断两组数据间的变异是否源自同一个原因。

15. Dynamic time warping

A method for locally stretching or compressing two one-dimensional vectors to correct for lag in one vector relative to another.

动态时间校正技术，能够将两种时间数据进行局部调整，使两者保持同步。

16. Mutual nearest neighbours (MNNs).

Cells that are mutually nearest to one another in normalized gene expression space.

彼此最靠近细胞，即在正常的基因表达空间里，彼此最靠近的细胞。

17. Cell-type classifications

Biologically meaningful labels given to groups of cells on the basis of common molecular profiles and prior knowledge of the cell types.

细胞分类，根据共同的分子特征和之前对细胞种类的认识，对细胞进行生物学分类

18. Gradient boosting

A statistical method that produces a prediction model for classification or regression on the basis of an ensemble of weaker prediction models.

梯度提升技术，这是一种统计学方法，可构建一个分类预测模型，也可对其它较弱的预测模型进行回归。

原文检索：

Tim Stuart and Rahul Satija.(2019)Integrative single-cell analysis. *Nature reviews Genetics*,

DOI: 10.1038/s41576-019-0093-7.

Eason/编译

特约编辑招聘启事

为了及时收集生命科学最新资讯、提高《生命奥秘》办刊质量，现面向从事生命科学或对这学科有浓厚兴趣的科研人员、学生诚聘特约编辑（兼职）。

岗位职责：

独立完成《生命奥秘》专题的策划：对基因组学、蛋白组学、生物信息学和细胞生物学等学科的发展以及生物医学领域相关技术（例如基因诊断技术、干细胞和克隆技术、生物芯片技术等）的应用进行翻译及深入评述。

选题要求内容新颖、评述精辟、注重时效和深入浅出。尤其欢迎以自身系统研究为基础的高水平译述与评论，结合所从事的科研工作提出自己的见解、今后设想或前瞻性展望。

要求：

1. 具备基因组学、蛋白组学、生物信息学、细胞生物学等生命科学学科背景；
2. 具备良好的生命科学前沿触觉；
3. 具备较高的外文文献翻译、编译水平；
4. 具备较强的选题策划、资料搜集、组织能力，以及专业稿件撰写能力；
5. 具有高级职称；或者拥有（正在攻读）该领域的最高学位。

有意者请将个人简历发送至 editor@lifeomics.com



单细胞转录组测序谱 & 基因功能研究产品和服务

系统化的优惠、快速服务 一起购买享**15%OFF**

1 10x Genomics单细胞转录组测序

- **高通量**：可进行1,000-10,000个单细胞转录组分析
- **高标准**：细胞标准品掺入，实时监控仪器及实验效率
- **好服务**：带仪器上门，完成10x单细胞建库
- **好团队**：曾参与863、973肿瘤研究项目，擅长数据挖掘，完成“测序谱”到“基因功能研究”

客户下单并提供样品信息

上门服务

单细胞悬液

细胞样品质检

10x Chromium
建库平台

文库质控

illumina
平台测序

数据分析

报告及可视化
结果

掺入细胞标准品

建库操作每一步，实时确定工作
流程性能标准！
有效提高实验准确率，保证成
功率！

资深分析团队，个性化定制分
析，满足客户不同的数据需求。

2 基因功能研究

- **多**：克隆类型多，如：启动子/3'UTR/shRNA克隆等；
提供超过 140,000 + 个ORF克隆；涵盖20000+个人
类基因；200+种哺乳动物、慢病毒等表达载体；
50+种蛋白融合标签
- **快**：Next-Day克隆，今天订货，明天发货
- **好**：克隆均保证序列，经过10+年全球科学家应用验证
- **省**：节省时间，价格优惠，克隆500元起



➤ miRNA研究方案

- miRNA克隆
- miRNA qPCR检测
- miRNA qPCR Arrays
- 报告基因检测
- miRNA 3'UTR克隆



➤ CRISPR基因组编辑

- 基因敲除
- 基因敲入
- 基因突变
- 功能验证



➤ 启动子研究

- 启动子报告克隆
- 分泌型双荧光素酶检测



➤ 分子生物学试剂

- RT & qPCR
- 荧光素酶检测试剂盒
- 转染试剂
- 支原体检测试剂盒
- 重组克隆体系 (LR&BP)
- NGS文库定量试剂
- Qubit® 定量检测试剂
- 荧光染料
- FISH荧光探针
- 感受态细胞
- DNA聚合酶



10x Genomics Chromium™ Single Cell 3' Solution 是一种单个细胞水平对mRNA进行高通量测序分析的高新技术。

从 = 测序 =

技术优势

- ① 不需要微量扩增，降低假阳性率；
- ② 真正的单细胞水平测序，Doublet rate : 0.9%/1 000个细胞；
- ③ 灵活的获取量，可一次性对500~80 000个细胞建库，提高效率；
- ④ 9分钟之内即可完成单细胞体系制备，捕获效率高达65%；
- ⑤ 广泛的研究应用领域，超短的项目周期。



应用领域

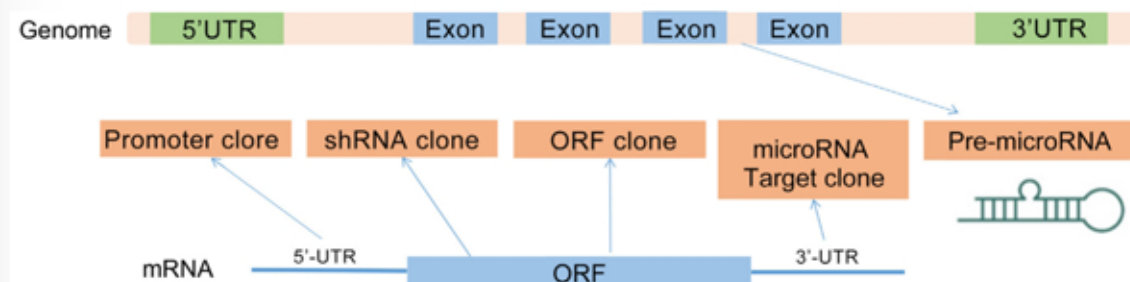
- ① 细胞群的分类研究；
- ② 大量细胞表达谱的检测；
- ③ 区分细胞类型组成或细胞状态的差异：如肿瘤细胞异质性，病理诊断，细胞周期；
- ④ 利用升级的单细胞RNA表达谱测序解析分子调控网络：可与CRISPR技术结合。

基因组编辑

- Cas9克隆/蛋白/细胞系
- HDR Donor 供体克隆
- sgRNA克隆
- Safe Harbor基因敲入
- IndelCheck检测体系

FISH检测

- VividFISH™ 染色体计数探针 (CEP)
- VividFISH™ 疾病相关基因探针 (LSI)



启动子研究

- 启动子克隆
- 荧光素酶检测试剂盒

ORF

- OmicsLink™即用型 ORF 表达克隆
- EZShuttle™ Gateway®PLUS 穿梭克隆
- 全长 cDNA 克隆

miRNA

- miRNA前体表达克隆
- miRNA抑制剂克隆
- miRNA 3' UTR 克隆

qPCR

- Gene/miRNA qPCR Arrays
- qPCR 试剂
- qPCR 验证引物

克隆构建服务

慢病毒/AAV包装服务

细胞鉴定及质控服务

细胞系构建服务

目的基因研究

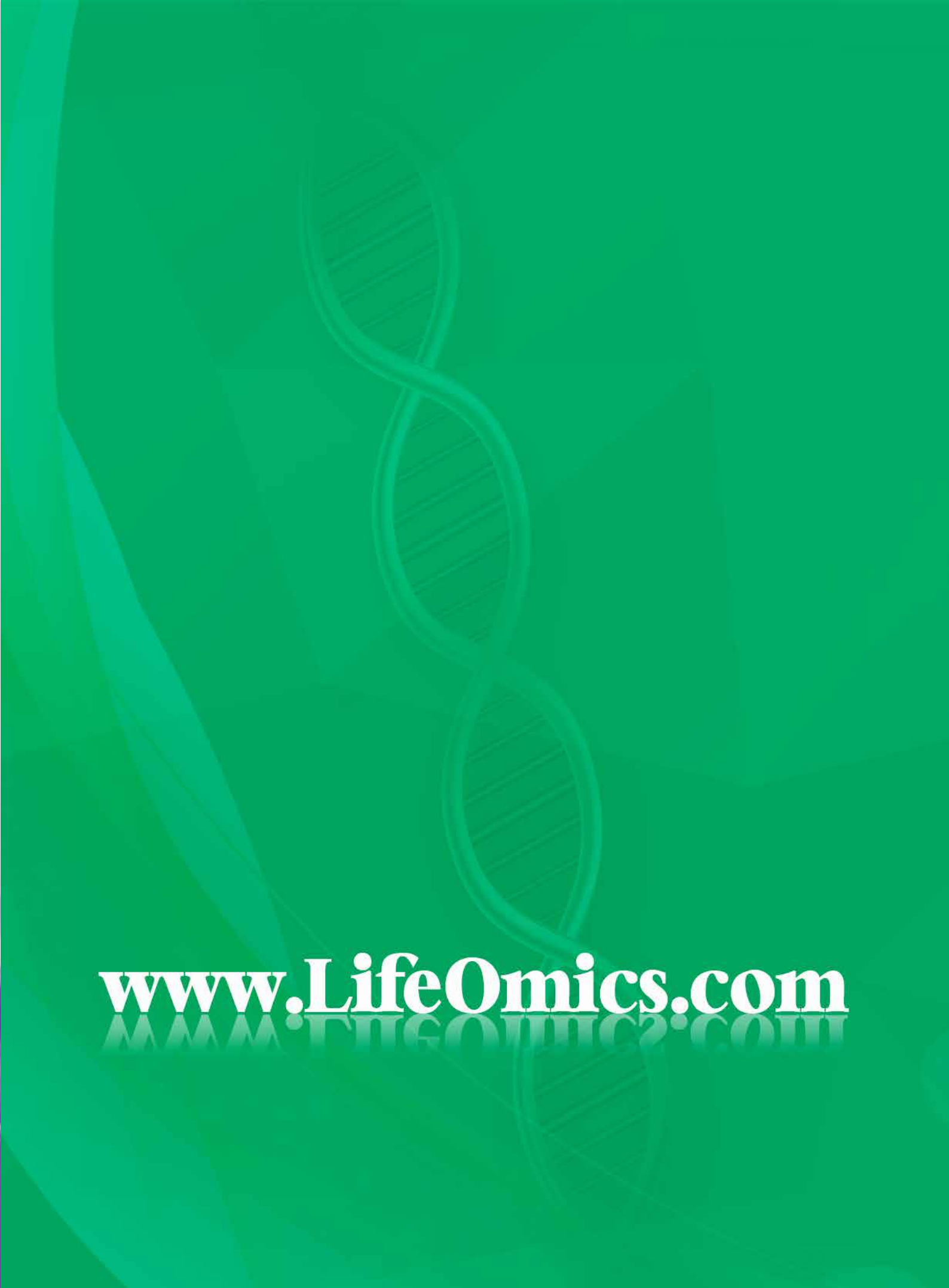


扫一扫，关注官方微信

A group of people are forming a human pyramid against a cloudy sky. The pyramid is composed of several layers of people, with the top layer having one person, the middle layer having two, and the base having three. The people are wearing dark clothing. The sky is filled with soft, white clouds. The overall tone of the image is motivational and collaborative.

合办专题专刊
网站广告合作
邮件群发推广

请致电 (020) 32051255



www.LifeOmic.com