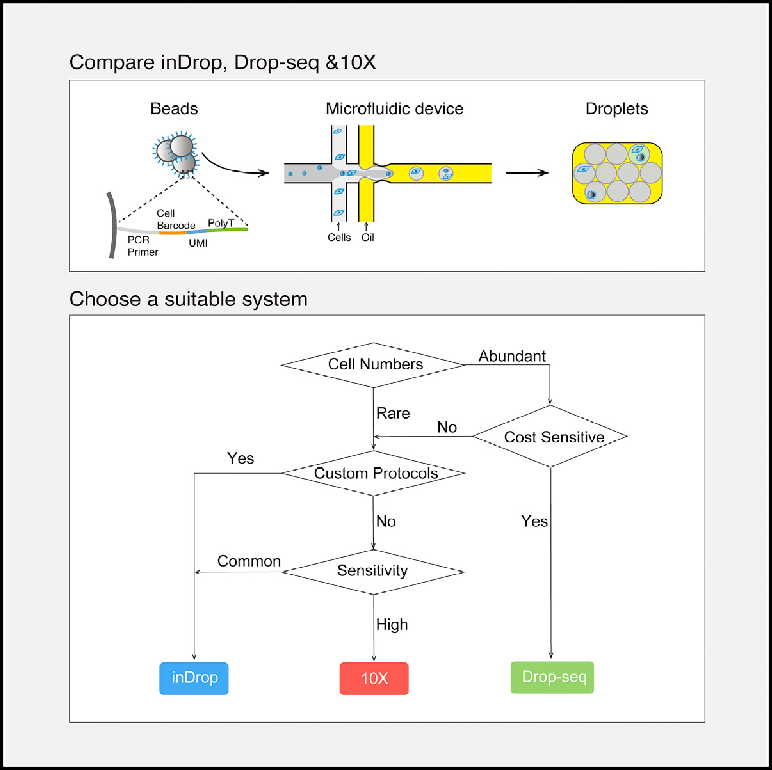
文章

基于液滴的超高通量单细胞RNA-Seq系统的比较分析

图形摘要



突出

d用于比较三个系统特点的综合图表

适用于所有系统的开源通用管道

d灵敏度、精确度、偏差和成本的系统比较

d使用inDrop平台演示Smart-seq2协议

作者

张先年，李天祁，柳峰，...、李泽耀、、黄、

一致[yanyi@pku.edu.cn](mailto:yanyi@pku.edu.cn)(Y.H .)，[jianbinwang@tsinghua.edu.cn](mailto:jianbinwang@tsinghua.edu.cn)(J.W .)

简单地说

Zhang等使用统一样本和生物信息学管道比较了三种流行基于液滴的高通量scRNA- seq系统。他们提供了系统设计和性能的详细分析，这将指导未来的实验设计和系统改进。

[](http://crossmark.crossref.org/dialog/?doi=10.1016/j.molcel.2018.10.020&domain=pdf)张等，2019，分子细胞73，130-142

2019年1月3日2018年Elsevier Inc .<https://doi.org/10.1016/j.molcel.2018.10.020>

分子细胞

文章

液滴基复合材料的对比分析

超高通量单细胞RNA-Seq系统

张先念，[1,](#_bookmark0)[6](#_bookmark5) 李天祁，[2,](#_bookmark1)[6](#_bookmark5) 柳峰，[3,](#_bookmark2)[6](#_bookmark5) 陈雅琪，[4,](#_bookmark3)[6](#_bookmark5) 姚家成，[2](#_bookmark1) 李泽耀，[5](#_bookmark4) 黄，[1,](#_bookmark0)[\*](#_bookmark7) 和王建斌[2,](#_bookmark1)[7,](#_bookmark6)\*

1北京基因组学高级创新中心(ICG)，学院生命科学学院生物医学创新创业中心(BIOPIC)

北京大学生命科学北京清华中心，北京100871

2清华大学生命科学学院和清华-北京生命科学中心，北京100084

3上海交通大学瑞金医院医学基因组国家重点实验室转化医学国家研究中心(上海)

童大学医学院，上海200025

4北京大学深圳研究生院，深圳518055Peking

5清华大学生命科学学院与北京-清华-美国国立卫生研究院联合研究生项目，北京100084School

6这些作者贡献均等These

7导联接触Lead

\*通信:[yanyi@pku.edu.cn](mailto:yanyi@pku.edu.cn)(Y.H .)，[jianbinwang@tsinghua.edu.cn](mailto:jianbinwang@tsinghua.edu.cn)(J.W .)<https://doi.org/10.1016/j.molcel.2018.10.020>

摘要

自2009年建立以来，单细胞RNA测序(RNA-seq)一直是生物医学研究进展的主要驱动力。在发育生物学和干细胞研究中，分析单细胞的能力具有特殊的优势。尽管大多数研究仍集中于单个组织或器官，但超高通量单细胞RNA-seq的最新发展已证明了其在表征更复杂系统甚至整个身体方面的潜在能力。然而，尽管多种超高通量单细胞RNA-seq系统已引起关注，但尚未对这些系统进行系统比较。在此，使用相同的细胞系和生物信息学管道，我们为三种广泛使用的基于液滴的超高通量单细胞RNA-seq系统inDrop、Drop-seq和10X Genomics Chromium分别开发了可直接比较的数据集。尽管每种系统都能够分析单细胞trans-scriptomes，但它们的详细比较揭示了每种系统的显著特征和合适的应用。

介绍

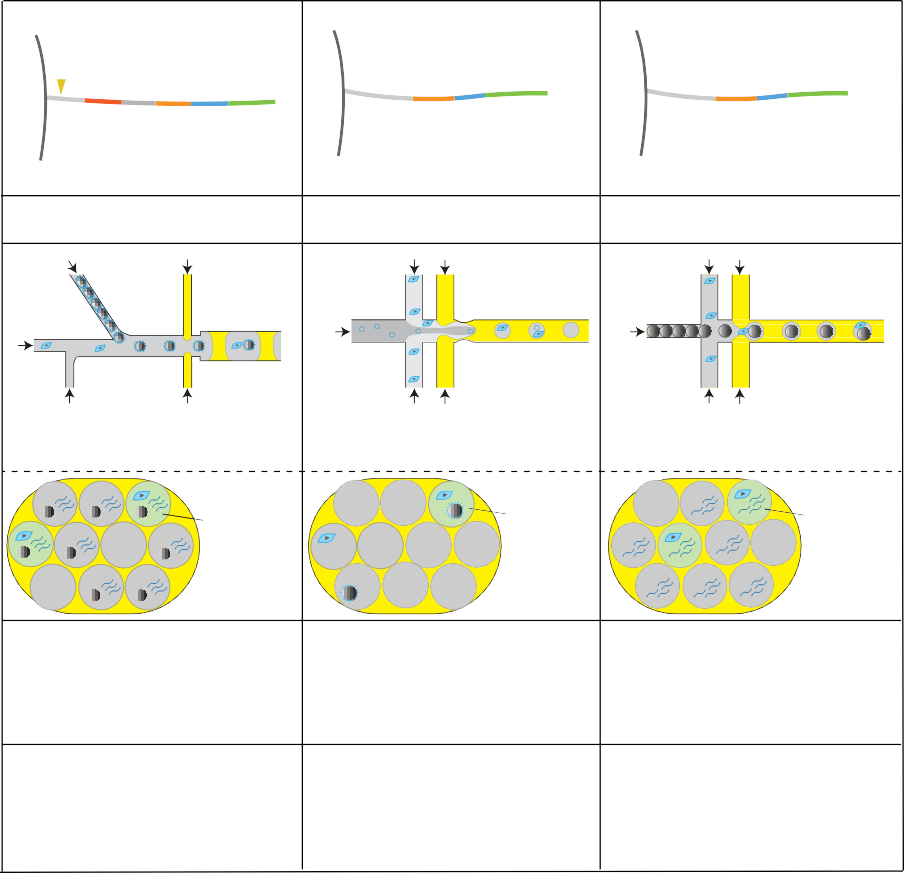
单细胞RNA测序(scRNA-seq)，最早于2009年建立([Tang et al., 2009](#_bookmark46))，已成为揭示生物异质性最有力的方法之一。在单细胞中操纵微微克RNA的能力使得研究能够以前所未有的速度和空间分辨率进行。基于整个转录组的大量数据，scRNA-seq以最精细的分辨率提供了关于基因表达及其调节相互作用的全面信息，从而能够准确和精确地描述细胞类型和状态([Gr](#_bookmark27)u[n and van Oude-](#_bookmark27)

[naarden, 2015; Tanay and Regev, 2017; Wu et al., 2017](#_bookmark27)).在过去的十年里，通过scRNA-seq进行mRNA定量的灵敏度和精确度都有了很大的提高([Hashimshonyet al., 2016; Picelli et al., 2014](#_bookmark29))，导致许多领域的革命性发现，例如在各种组织或器官中的细胞类型鉴定([Jaitin et al., 2014; Lake et al., 2016; Papalexiand Satija, 2018; Treutlein et al., 2014; Villani et al., 2017](#_bookmark31));追踪胚胎发育和细胞分化中的细胞谱系和命运承诺([Olsson et al., 2016; Semrau et al.,2017; Tirosh et al., 2016; Yan et al., 2013](#_bookmark35));对转录动力学和调节网络进行推断([Denget al., 2014; Dixit et al., 2016](#_bookmark22));以及鉴定肿瘤的发展、演变和异质性([Patel et al., 2014; Treut-lein et al., 2014; Venteicher et al., 2017](#_bookmark37)).

实验吞吐量一直是scRNA-seq实验设计中的一个主要关注点。在一些生物系统中，例如早期胚胎，仅需要几十个细胞来实现关键发现([Yan et al., 2013](#_bookmark47)).然而，由于组织的复杂性、细胞周期的动态性或其它内在变化([Buettner et al., 2015](#_bookmark20))，以及技术噪音([Brennecke et al., 2013](#_bookmark18))，来自少量细胞的RNA-seq数据通常不足以全面反映生物样品的状态([Tanay and Regev, 2017](#_bookmark45)).已知转录组检测的灵敏度随着测序深度的增加而迅速饱和([Svenssonet al., 2017](#_bookmark44)).大规模采样单细胞的浅层测序可以有效减少随机变异，并通过聚类分析定义细胞类型，提供了一种更为稳健的方法([Pollen et al., 2014; Streets and Huang, 2014; Svens-son et al., 2018](#_bookmark38)).对于大规模的scRNA-seq研究，一个主要的技术障碍是制备大量cDNA文库的成本。实验室自动化可以克服这种方法的劳动强度，但是试剂仍然昂贵([Jaitin et al., 2014](#_bookmark31)).最近报道的一些微流体方法已经证明了scRNA- seq([Prakadan et al., 2017](#_bookmark39)).例如，当与适当的化学物质结合时，小体积反应器可以提高反应效率并降低技术噪音([Streets et al., 2014;Wu et al., 2014](#_bookmark43)).此外，芯片实验室操作也使单细胞分离比手动挑选细胞容易得多

[](http://crossmark.crossref.org/dialog/?doi=10.1016/j.molcel.2018.10.020&domain=pdf)130 Molecular Cell 73，130–142，2019年1月3日2018 Elsevier Inc .

1. inDrop Drop-seq 10X



珠子

油

细胞

油

试剂细胞

油

0.1小时

细胞

珠子

裂解试剂

珠子

逆转录/裂解试剂

油

细胞

油

细胞试剂

油

珠子:超中毒细胞:中毒细胞

珠子:泊松细胞:泊松

珠子:超中毒细胞:中毒细胞

7小时

* 光电导继电器(Photo-conductive Relay)
* cDNA片段化和连接
* 光电导继电器(Photo-conductive Relay)
* 逆转录酶和模板开关
* 光电导继电器(Photo-conductive Relay)
* Tn5标记
* 光电导继电器(Photo-conductive Relay)

9小时

* 第二链合成
* 体外转录
* RNA片段化
* 逆转录聚合酶链反应

28小时

1小时

* 细胞溶菌作用
* 通过珠溶解释放引物
* 逆转录和模板转换

0.3小时

* 细胞溶菌作用
* 磁珠上的mRNA捕获

2.5小时

* 细胞溶菌作用
* 通过紫外线释放底漆
* mRNA捕获
* 反转录

裂解/反应混合物

裂解混合物

裂解/反应混合物

0.5小时

734,000

16,777,216 (412)

147，456 (384 X 384)

凝胶

UMI

10bp

多面体3 '

16bp

条形码

细胞

5'

光电导继电器(Photo-conductive Relay)

底漆

UMI

8bp

多面体3 '

12bp

条形码

细胞

5'

光电导继电器(Photo-conductive Relay)

底漆

水凝胶

3'

UMI

6bp

光电导继电器(Photo-conductive Relay)

底漆

5'

多面体

38~41bp

条形码

细胞

可光裂解的接头

T7p

0.3小时

条形码引物珠

可变形的

困难的

可分解的

细胞条形码

容量

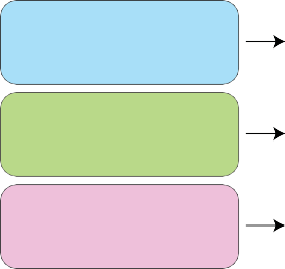
液滴产生

乳状液

液滴中的反应

破乳后的反应

1. 输入单元格数据处理



**inDrop运行1: ~8，300个细胞**

运行2: ~8，300个单元格

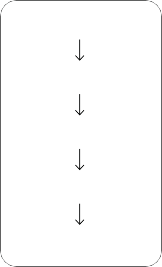
运行1: ~100，000个单元格

**Drop-seq Run 2: ~ 50，000个单元格**

运行3: ~ 50，000个细胞

**10X**

运行1: ~ 3，500个单元格运行2: ~14，000个单元格



条形码检测

对齐

基因标签

UMI计数

细胞过滤

已识别的细胞

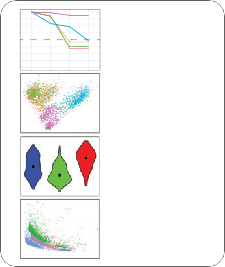
比较





RNA UMI CB

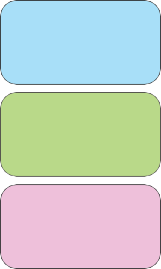


读取内容

方法偏差

敏感性

技术噪声



运行1: 2，235个单元格

运行2: 1，883个单元格

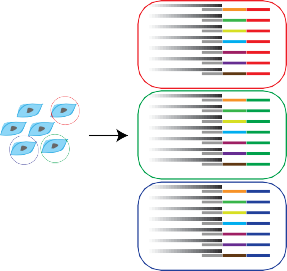
运行1: 2，240个单元格

运行2: 1，157个单元格

运行3: 1，179个单元格

运行1: 1，560个单元格

运行2: 6，478个单元格



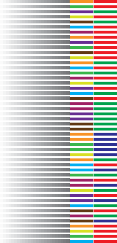
细胞

**C**

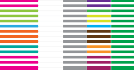
...

...



基因





单元格#1

一

2

2

一

一

...

2号牢房

2

一

一

一

2

...

3号牢房

2

一

四

...

...

条形码/标签测序解复用UMI计数

...

...

...

...

*(下一页图例)*

分子细胞73，130-142，2019年1月3日131

（[Shalek et al., 2014](#_bookmark42)).基于微孔的scRNA-seq方法([Fanet al., 2015; Han et al., 2018](#_bookmark25))在低成本和高通量方面也显示出优势。然而，由于缺乏市售仪器或详细方案，基于微孔的scRNA-seq未被广泛采用。

液滴微流体可以在高达每秒数十万个液滴的频率下实现快速分隔和封装，并且可以容易地按比例缩放以产生数百万个液滴，每个液滴具有纳升体积以适应单细胞反应([Agresti et al., 2010](#_bookmark16)).微流体管道布局非常简单，主要由引入或收集试剂和样本的微通道组成([Duncombe et al.,2015](#_bookmark24)).这种液滴策略极大地提高了反应吞吐量，并显著降低了成本。目前，用于高通量scRNA-seq的基于液滴的系统有三种，即inDrop([Briggs et al., 2018; Klein et al.,2015; Wagner et al., 2018; Zilionis et al., 2017](#_bookmark19))，拖放序列([Farrellet al., 2018; Macosko et al., 2015](#_bookmark26))，以及10X基因组铬(10X)()[Zheng et al., 2017](#_bookmark48)).所有这些都已被证明在以可接受的成本一次性生成数千个细胞的cDNA文库方面是稳健和实用的。所有三种方法都使用类似的设计来生成液滴，使用带有条形码的珠上引物来区分单个细胞，并应用唯一的分子标识符(UMI)进行偏差校正([Kivioja et al., 2011](#_bookmark32)).尽管存在这些相似性，但它们涉及不同的磁珠制备、条形码设计和cDNA扩增方法，因此具有不同的实验方案。鉴于系统规格和转录组分析结果的这些差异([Ziegen-hain et al., 2017](#_bookmark49))，需要对这些方法进行系统和无偏倚的比较。

这里，我们使用同一个样本和一个统一的数据处理管道来比较这三种方法的性能。对于每种方法，我们使用淋巴母细胞细胞系GM12891进行了2至3次复制。每个细胞条形码的平均测序深度约为50，000次读取。我们还开发了一个多功能、快速的数据处理工作流程，并将其应用于所有数据集。分析比较细胞捕获效率、有效读取比例、细胞条形码错误率、转录物检测灵敏度。研究结果揭示了各系统的优缺点，为未来研究中选择最合适的系统提供了指导。

结果

系统概况

在这三种系统中，inDrop和Drop-seq已在文献中进行了广泛描述，而10X是一种商业平台，其设计细节尚未完全公开。

我们在此试图根据我们所能收集的信息，尽我们所能剖析这些系统。在所有三种系统中，细胞条形码嵌入微珠-栓系引物([Figure 1](#_bookmark8)a)。珠上引物的DNA序列具有共同的结构，包括PCR手柄、细胞条形码、UMI和聚inDrop珠上的引物还具有可光裂解部分和T7启动子。然而，珠子是用不同的材料制成的。10X和inDrop系统中使用的磁珠由水凝胶制成，Drop-seq使用脆性树脂。通常，以低浓度引入珠子和细胞以减少形成双链体的机会；也就是说，两个细胞或两个珠子被包封在单个液滴中。因此，对于使用小硬珠的液滴序列，同一液滴中一个珠和一个细胞的封装遵循双泊松分布。水凝胶珠是软的且可变形的，紧密地封装在微流体通道中，并且它们的封装可以同步以实现超泊松分布([Figure 1](#_bookmark8)a；[Abate et al., 2009](#_bookmark15)).尽管由于珠大小不可避免的变化，100%单珠占有非常困难，但在10X和inDrop方法中，细胞捕获效率可以达到明显更高的水平。据报道，10X具有~80%的珠占有率和~50%的细胞捕获率([Zheng et al., 2017](#_bookmark48)).

珠子的材料也可能影响DNA引物的数量和密度。对于10X和inDrop使用水凝胶允许在整个珠子中固定引物，而较小的Drop-seq珠子只能在表面携带引物。封装后，从10X开始的整个磁珠被溶解，以将所有引物释放到溶液相中，从而提高mRNA捕获的效率。inDrop还通过紫外线辐射诱导的裂解来动员引物。相比之下，Drop-seq使用表面拴系引物来捕获mRNA分子，这可能会降低捕获效率(与10X和inDrop相比)。

反转录在破乳前对10X和inDrop的液滴进行。相反，Drop-seq仅捕获转录物，而不进行cDNA转化。由于分离了许多局部反应并减少了反应竞争，液滴中的逆转录可以获得更一致的结果。还已知在有限体积(例如液滴)中的反应性能提高了cDNA转化的特异性和相对产率([Streets et al., 2014](#_bookmark43)).这三个系统采用不同的cDNA扩增策略。inDrop采用CEL-seq([Hashimshony et al., 2012](#_bookmark28))，而10X和Drop-seq遵循模板切换协议([Macosko et al., 2015;Zheng et al., 2017](#_bookmark36))类似于流行的Smart-seq化学([Ramsko¨ ld et al., 2012](#_bookmark40)).inDrop中的体外转录步骤延长了库制备时间，超过24小时，尽管Drop-seq和10X管道都可以在一天内完成。

图1。三大平台、实验设计和数据分析管道概述

1. 三种系统的实验特点示意图及比较。它们在条形码设计、库大小、乳剂和下游反应方面存在差异。
2. 实验方案总结。对每个平台进行了两次或三次重复，并使用相同的数据处理管道进行下游分析。对输入和恢复的细胞编号进行了标记。
3. 数据处理管道工作流概述。由逆转录中的条形码和标签产生的测序读数首先被它们的细胞条形码解复用，然后被映射到每个基因的umi被聚集和计数。

另见图S1。

132分子细胞73，130-142，2019年1月3日

实验设计与数据处理

我们使用人淋巴母细胞样细胞系GM12891进行比较研究。为所有三个系统设置了生物复制，在不同的日期和不同的批次中输入不同的细胞数([Figure 1](#_bookmark8)b)。我们调整了测序深度，以获得三个系统中每个细胞条形码的可比读取数(参见[STAR Methods](#_bookmark50)).

每个系统都有自己的数据处理管道。然而，由于读取结构的不同，它们都不能直接处理其他系统产生的数据。每个分析流水线都必须处理依赖于系统的数据特性，例如细胞条形码误差的容限。此外，分析管道在一些关键过程中使用不同的策略，如基因标记。所有这些差异都可能在基因定量中引入偏差，这在试图对系统进行公平比较时并不理想。为了解决这个问题，我们开发了一个多功能管道，它接受来自所有这些系统的数据，并生成UMI计数矩阵([Figure 1](#_bookmark8)c)。我们将此管道应用于我们的数据，并以客观的方式对灵敏度、精密度和偏倚进行了比较。

管道包可在线免费获得(<https://github.com/beiseq/baseqDrops>)下载。设计用于接受成对端测序数据，一端(读1)包含细胞条形码和UMI，另一端(读2)包含转录物序列。流水线首先识别read 1原始数据中的细胞条形码。移除读取计数过低的细胞条形码(杂项条形码)后，管道会更正细胞条形码错误(请参阅[STAR Methods](#_bookmark50)).这些错误可能是在珠上引物合成过程中以及PCR或测序步骤中引入的。对来自相同细胞条形码的读取进行聚合，并在按读取计数过滤后移除无效细胞条形码。对于条形码不完全随机的10倍和inDrop，该管道根据制造商的白名单进一步过滤细胞条形码。

使用STAR将读取2序列映射至人类参考基因组(hg38)()[Dobin et al., 2013](#_bookmark23))然后标记到相应的基因上。我们还使用每个协议的官方管道处理数据集。然后，我们将获得的结果与我们多功能管道的结果进行了比较。发现大多数基因的表达水平和每个条形码中的UMI计数在不同的数据处理方法之间高度一致(参见[STAR Methods](#_bookmark50)；图S2A和S2B)。为了确认将对齐的读数转化为相应基因的准确性，我们通过基于细胞系的基因表达谱产生约200万个读数来进行模拟(参见[STAR Methods](#_bookmark50)).超过99%的读数(2，251，529个读数中的2，229，156个)被标记到正确的基因上(参见[STAR Methods](#_bookmark50)；图S2C)。其余1%的歧义阅读主要来自具有旁系同源物或重叠基因的基因，如RPL41/AC090498.1或IGHA1/ IGHA2(表S2)。在读-基因分配后，对每个细胞中每个基因的读数进行分组，通过允许1-bp的错配对其UMI进行聚集和计数，从而生成基因表达UMI矩阵。

该流水线的处理速度通过减少读写有效负载得到了优化，读写有效负载是常见的瓶颈-

脖子。例如，大约50%的inDrop数据读取具有无效的序列结构。通过删除这些读取，我们可以提高数据处理效率。此外，基于单元条形码前缀，读取被分成多个(通常N = 16)文件，从而实现并行处理。

磁珠上引物的质量

条形码库的大小决定了使用基于液滴的scRNA-seq进行单次实验运行的最大容量。一个小的细胞条形码库可能会导致条形码碰撞和人为的双重。在三个系统随附的信息中，理论细胞条形码库大小为1.47 3 105 (in-

Drop)、1.6 3 107 (Drop-seq)和7.34 3 105 (10X)被要求赔偿。

但是，有效条形码库大小可能小于设计值。我们通过分析每个系统多次运行之间的条形码冲突来估计有效条形码的比例(请参阅[STAR Methods](#_bookmark50)).可能性分析证明了在不同的有效条形码分数下观察到这样数量的碰撞的相对概率([Fig-ure 2](#_bookmark9)a)。对于inDrop，我们的结果表明有效的条形码前体部分约为30%，尽管100%的有效性也是可能的，且可能性较小。对于较大的库，分析的功能较弱，但我们仍可尝试确定Drop-seq (~10%)和10X (~40%)的有效比例下限。有效条形码比例小于下限的可能性相对较低。因此，粗略估计，inDrop的有效条形码大小为5 3 104，至少为1 3

~

Drop-seq为106，10X为3 3 105(参见[STAR Methods](#_bookmark50)).

一个条形码一个磁珠是所有三种系统的关键要求。然而，由于DNA合成化学的不完善，异步碱基添加是不可避免的。因此，同一磁珠内可能出现细胞条形码序列不一致的情况。细胞条形码中出现此类错误将导致检测到的单细胞数量激增，这需要仔细校正。我们在1汉明距离内聚集了细胞条形码。对于每个有效的细胞条形码，校正后的读数(包含原始条形码序列中的错误)占校正后总读数的比例计算为细胞条形码错误率([Figure 2](#_bookmark9)b)，反映了珠上DNA引物的一般质量。10X磁珠的细胞条形码几乎没有错配，表明磁珠制备过程中质量控制良好。相比之下，在其他两个系统中，超过一半的细胞条形码包含明显的错配。具体而言，约10%的Drop-seq珠子在细胞条形码中含有一个碱基缺失，这在数据分析期间也需要格外小心(参见[STAR Methods](#_bookmark50)).

我们进一步分析了UMI的碱基组成，这可以反映它的合成和使用偏差([Figure 2](#_bookmark9)c；表S1)。所有系统都表现出对poly-T的偏好或偏爱，因为它与mRNA的poly-A尾部具有亲和力。我们还发现inDrop中的poly- C富集，Drop-seq和10X中的poly-G富集。这种模式，主要是由于DNA合成偏差，可能导致RNA-seq结果的系统依赖性偏斜。

有效细胞条形码的主要过滤标准是基于原始读取的总数，这在很大程度上反映了细胞mRNAs的丰度。读取次数更多的细胞条形码更可能源自真实细胞。通过读取计数对细胞条形码进行分类和可视化，我们观察到

分子细胞73，130-142，2019年1月3日133

# A

0.04

InDrop Drop-seq 10X

0.03

可能性

0.02

0.01

0

0.25 0.50 0.75 1.00

0.6

0.4

0.2

0

0.25 0.50 0.75 1.00

有效条码比例

0.075

0.050

0.025

0

0.25 0.50 0.75 1.00

# B

100

75

百分率

50

25

0

细胞条形码错误率

>0.5

0.25-0.5

0.05-0.25

<0.05

**C 1.0**

0.5

0.0 GG

2.0

拖放序列

游戏结束

A

电汇

常规军备委员会(Committee for Conventional Armaments)

游戏结束

TTT·T

游戏结束

游戏结束

T

助教A A A A A A

C

位

C

|  |  |
| --- | --- |
| InDrop | |
|  | CCTT  TCC |
| 在TAA大学  T | |

1.0

位

0.0

2.0

10X

G 集料

C

电汇

T 气相色谱

A A GG

A

TTG

T

T

T

GTG

A A

C

5

10

1.0

位

0.0

G

G

游戏结束

5

5

# D

1e+05

1e+04

1e+03

1e+05

读取计数

1e+04

1e+03

1e+05

1e+04

1e+03

inDrop-1 inDrop-2

Drop-seq-1 Drop-seq-2 Drop-seq-3

10X-1

10X-2

10 100 1000 10000

条形码位置条形码等级

图2。珠上引物库大小和质量评估

1. 估算每个系统的有效细胞条形码库大小。显示了不同有效条形码比例的可能性。可能性分析基于观察到的来自同一系统的不同样本之间的条形码冲突(请参阅[STAR Methods](#_bookmark50)).
2. 细胞条形码错误率的分布。错误率是以更正读数(1-bp不匹配)相对于总读数的比例来衡量的。
3. 每个系统最常用的50个umi的主题。
4. 根据读取计数对有效细胞条形码编号的初步估计。同一样本中的细胞条形码按其读取计数排序。根据输入的细胞数量和实验捕获效率选择前N个细胞条形码。

另见表S1。

三个系统中的功能([Figure 2](#_bookmark9)d)。对于10倍，陡崖表明来自健康细胞的barco- des与其他细胞的barco-des之间的读取计数存在明显差异。对于inDrop来说，有一个相似但更微妙的悬崖。然而，对于Drop-seq，在读取计数曲线上没有明显的悬崖以获得清晰的界限。这可能源于Drop-seq使用的珠子的宽尺寸分布。我们注意到inDrop或10X中使用的珠子的尺寸比Drop-seq中使用的珠子的尺寸更均匀(图S1)，这可能是由于制造树脂珠子时尺寸控制的困难。

数据处理步骤和结果

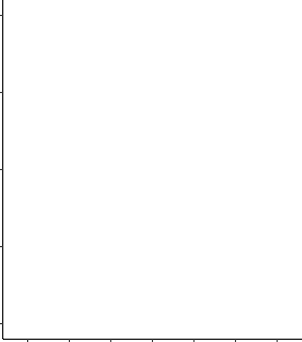
准确测定每个样本中由细胞条形码表示的细胞数量具有挑战性。这是由于细胞mRNA分子计数的大分散及其捕获效率所致。我们尝试了多种策略来估计有效的细胞数量(见[STAR Methods](#_bookmark50)；图S3)。其中许多方法依赖于对读数或umi分布或细胞组成的某些假设，这些假设可能并不适用于所有方案或情况。我们实施了一种策略，从实验确定的给定数量的细胞开始，然后进行严格的质量控制过滤(UMIs R 1，000和最近的correla-

R 0.6)。在最近报道的高通量scRNA-seq研究中，多组已经实施了这一策略。对于每次运行，可根据系统特定方案，通过考虑输入细胞的数量、细胞捕获率和下游反应成功率，粗略估计回收细胞的数量。然后，进一步过滤估计的细胞以满足质量控制标准(参见[STARMethods](#_bookmark50)).

分裂成每个有效细胞条形码的读数首先与人类基因组对齐以分析整个基因组中读数的分布([Figure 3](#_bookmark10)a)。Drop-seq和10X约有65%的读数被映射到UTR(主要是30个UTR)和外显子区域，尽管这一比例在inDrop中仅为约45%。在对映射到基因体的读数进行标记后，可获得可检测基因的数目([Figure 3](#_bookmark10)b)。10X-1以及Drop-seq-2和Drop-seq-3中的基因数量下降是由于输入细胞数量很少。除Drop-seq数据中的几个异常值外，基因数量根据细胞内读取的数量而下降。我们使用这些检测到的基因来证明沿着基因体的读分布的偏差([Figure 3](#_bookmark10)c)。所有三个系统的读数主要来自mRNA的30端，与

134分子细胞73，130-142，2019年1月3日

# A



100

B 8000

75

百分率

未在地图上标明的

50个基因间

内含子编码

25 UTR

0

6000

4000

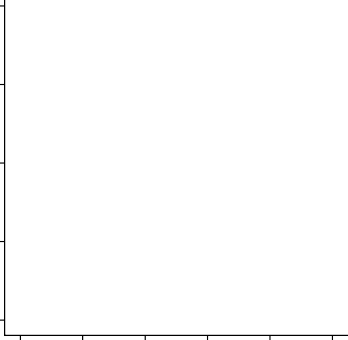
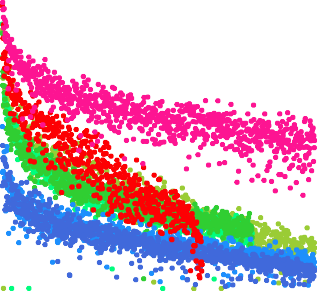
基因

2000

0

0

500



1000 1500 2000 2500

条形码等级

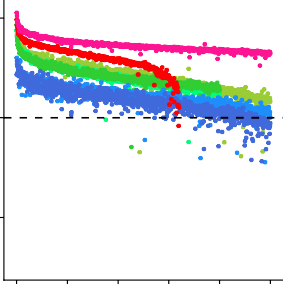
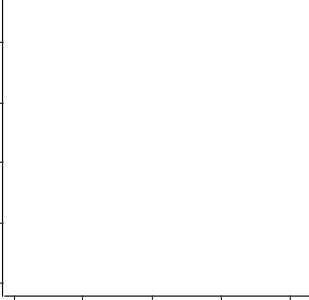
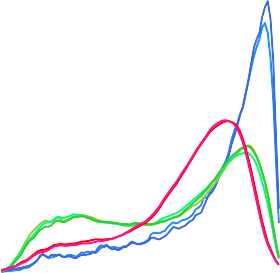
inDrop-1 inDrop-2 Drop-seq-1

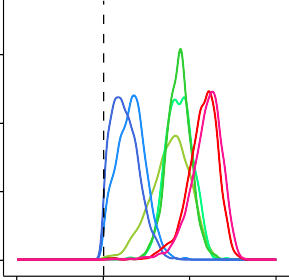
Drop-seq-2

放下seq-3 10X-1

10X-2

# D E



4 1e+05

15

3

密度

密度

1e+03 10

UMIs

2

1 1e+01

5

0

0 25 50 75 100

标准化基本位置(%)

0 500 1000 1500 2000 2500

条形码等级

0

0.4

0.6 0.8

最近相关

1.0

# F G



样本估计UMI 1K最近相关

有效阅读比例

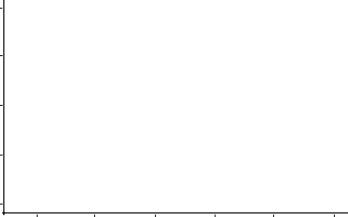
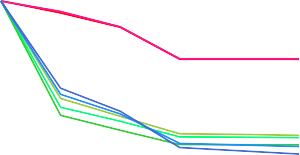
>=

>= 0.6

|  |  |  |  |
| --- | --- | --- | --- |
| inDrop-1 | 2,500 | 2,447 | 2,235 |
| inDrop-2 | 2,500 | 2,248 | 1,883 |
| Drop-seq-1 | 2,500 | 2,430 | 2,240 |
| Drop-seq-2 | 1,200 | 1,161 | 1,157 |
| Drop-seq-3 | 1,200 | 1,181 | 1,179 |
| 10X-1 | 2,000 | 1,567 | 1,560 |
| 10X-2 | 8,000 | 6,483 | 6,478 |

1.00

0.75



0.50

0.25

0

图3。数据处理步骤和结果

1. 在映射到基因组后阅读作文。显示了映射到不同基因组区域的读取和未映射读取的百分比。
2. 使用细胞条形码检测到的基因数按读取计数排序。
3. 基因体从50到30端的标准化阅读分布。
4. 具有按读取计数排序的细胞条形码的umi数量。
5. 细胞与同一样本中所有其他细胞的最近相关性分布(参见[STAR Methods](#_bookmark50));将0.6的阈值应用于质量控制。
6. 在质量控制筛选的每个步骤之后保留的有效细胞条形码的数量。
7. 质量控制过程每一步后的有效读数比例(参见[STAR Methods](#_bookmark50)).另见图S2和S3。

他们的图书馆建设策略。Drop-seq数据显示了双峰分布，这很可能是由于cDNA分子两端使用了相同的PCR锚定序列。

我们基于每次实验运行中的umi(转录物)总计数进行细胞条形码过滤([Figure 3](#_bookmark10)d)。带着微笑

总UMI截止值为1000时，大部分细胞条形码都通过了筛选，这表明估计的细胞数量是正确的。为了进一步去除由条形码错误引起的可能伪影，我们检查了相似细胞条形码之间表达谱的相似性。如果细胞条形码的表达谱是

分子细胞73，130-142，2019年1月3日135

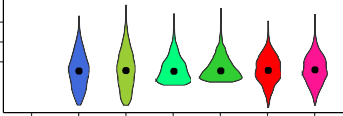
# 甲和乙



样本细胞条形码Raw读取UMI读取Genes UMIs

2e+05

1e+05



读

5e+04



inDrop-1

2,122 36,876 19,693

1,255

2,824

inDrop-2 1，636 36，338 17，509 1，241 2，645



Drop-seq-1

2,240 36,796 24,314

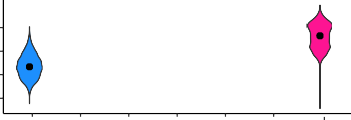
2,196

6,923

Drop-seq-2 1，154 36，204 24，329 2，590 8，532

4000

2000

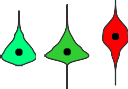
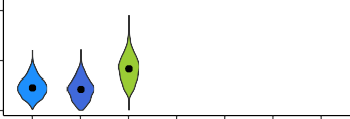


基因

1000

500

1e+05



Drop-seq-3

1,178 36,512 23,612

2,638

8,421

10X-1 1，558 37，136 26，796 2，996 17，026

1e+04

1e+03



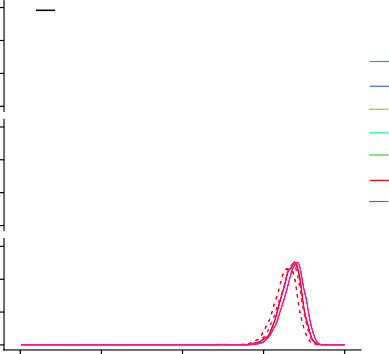
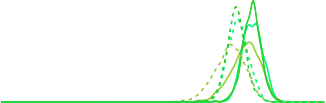
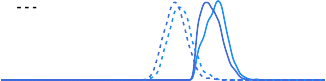
10X-2

6,476 36,713 26,144

3,107 16,828

UMIs

# C



UMI使用读取计数

15

10

5

0

15

10

密度

5

0

15

10

5

0

0.00 0.25 0.50 0.75 1.00

最近相关

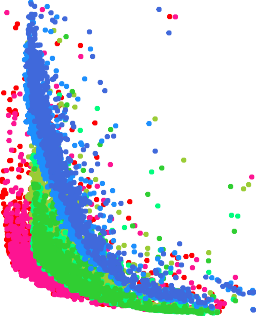
inDrop-1 inDrop-2

Drop-seq-1 Drop-seq-2 Drop-seq-3 10X-1

10X-2

**D**

四



3

2

CV2



一

0



1e+02 1e+03 1e+04 1e+05

每百万人的平均umi

inDrop 1 inDrop 2

Drop-seq 1 Drop-seq 2 Drop-seq 3 10X 1

10X-2

图4。各平台的灵敏度和技术噪声演示

1. 细胞条形码编号、读取计数和分子检测性能总结。对数据进行下采样，以在所有样本中获得一致的原始读取水平(参见[STAR Methods](#_bookmark50)).
2. 检测到的原始读数、umi和基因的分布。
3. 技术噪声通过同一样本中一个细胞条形码与其他每个细胞条形码之间的最近相关性来测量。通过UMI计数(实线)和读取计数(虚线)的基因定量都被采用。
4. 每个系统的CV-平均值(CV平方)图。技术噪声是在基因水平上测量的。另见图S4和S5以及表S3和S5。

明显不同于其最接近的细胞条形码邻居(Spear- man相关性% 0.6；看见[STAR Methods](#_bookmark50))，我们丢弃了条形码([Figure 3](#_bookmark10)e；看见[STAR Methods](#_bookmark50)).

通过所有这些步骤，我们在每个实验中获得了不同数量的细胞([Figure 3](#_bookmark10)f)。有效读取(从有效条形码读取)的比例对于10X为~75%，对于inDrop为~25%，对于Drop-seq([Figure 3](#_bookmark10)g)。此类读取的比例应最大化，以减少测序容量的浪费。

UMI和基因检测的灵敏度基因检测的灵敏度是scRNA-seq性能的一个基本指标。它反映了捕获单个mRNA分子用于反向转录、第二链合成和预扩增的方法的总体效率。它进一步影响和决定基因表达定量的精度和准确度。与一个

输入样本，灵敏度可以简单地用回收的umi和基因计数来描述([Figure 4](#_bookmark11)a)。随着读取计数的增加，细胞条形码的UMI数和基因数逐渐饱和(图S4A和S4B)。我们发现，经过对数转换的UMI计数与检测到的基因数量高度相关(Spear- man相关r > 0.9)(图S4C)。这表明测序深度可能会影响umi和检测基因的数量。为了在三个不同系统之间进行公平比较，我们对数据集进行了标准化，以在基因表达分析之前达到统一的原始读取水平(~36K/cell)(参见[STAR Methods](#_bookmark50)).来自同一系统的技术回复显示出高一致性和再现性。10X的灵敏度最高，平均捕获约3，000个基因的17，000多个转录本。这种性能是一致的，与输入单元的数量无关。Drop-seq检测到约2，500个基因的约8，000个转录物。同时，inDrop系统的检测灵敏度较低

136分子细胞73，130-142，2019年1月3日

来自约1，250个基因的约2，700个umi。inDrop和Drop-seq中的读取分布更为偏斜，对于这两种情况，大多数细胞条形码的读取计数相对较低([Figure 4](#_bookmark11)b)。

技术噪声和精度

技术噪声反映了实验随机性所赋予的变化，包括逆转录过程中的转录物丢失以及与PCR扩增相关的偏倚。精确度可以通过技术复制品之间转录组的一致性来评估。执行单细胞RNA-seq的主要目的是根据细胞的基因表达谱将细胞聚类成不同的亚组，通常是为了发现和表征新的细胞类型或状态。聚类是基于细胞间基因表达模式的相似性或距离。较大的技术噪声或变化会扭曲实际距离，掩盖细胞之间微妙的生物学差异，从而降低细胞分组的分辨率。为降低技术噪声，人们做了很多努力，如利用UMI消除放大偏差引起的量化误差。

虽然我们在这里使用了一个明显同质的细胞系，但仍然存在内在的生物噪声或异质性([Prakadanet al., 2017](#_bookmark39)).在我们的数据集中，总变异由技术和生物成分组成，很难分开。这里，我们假设生物噪声在样本之间是一致的，并且技术噪声主导了数据集的变化。看家基因(显示最低水平的生物噪声)和其他基因的噪声水平具有相似的分布，这表明与技术噪声相比，生物噪声水平较低(图S5；看见[STAR Methods](#_bookmark50)).因此，总体总变化应反映技术噪声水平。

总变异量的特征是整个数据集中特定细胞条形码与每隔一个细胞条形码之间最接近的Spearman相关性(参见[STARMethods](#_bookmark50)).许多聚类或分类策略，如k-means和分层聚类，都是基于细胞之间的最近相关性。要识别次要细胞类型，这些次要细胞之间的最近相关性应较高，以使其能够与其他细胞分离。为了验证UMI在降低基因计数的PCR扩增噪声方面的效果，我们使用UMI计数和原始读数计数进行分析，以量化基因表达。结果([Figure 4](#_bookmark11)c)表明10X和Drop-seq的技术噪声水平低于inDrop。对于所有三个系统，以UMI为特征的基因表达谱与使用原始计数的基因表达谱相比降低了噪声，从而证实了UMI在降低噪声方面的有效性。值得注意的是，这种噪声在inDrop数据中更为严重，可能是由于在文库构建过程中使用了随机引物。然而，对于10倍，UMI的使用并没有显著降低噪音。这可能是由于在10X样本制备过程中扩增相对均匀所致。此外，大多数umi仅测序两到三次，表明测序深度不够饱和。对于更深层次的测序，使用UMI可能会进一步降低噪音。

基因水平的技术变异可以通过所有细胞的标准化UMI(umi/百万)计数的变异系数(CV)来测量([Figure 4](#_bookmark11)d；看见[STAR Methods](#_bookmark50)).这为整个基因的技术噪音提供了一个视角

表达式配置文件。对于表达水平较高的基因，所有系统的变异均减少。通常，10X的技术噪声最低，其次是Drop-seq，然后是inDrop。有趣的是，许多高表达的基因相当嘈杂，尤其是在10倍的数据中。我们检查了这些基因(正常化的UMI R 2，000；CV R 0.5)，发现其中大部分是该细胞系的最高表达基因或线粒体基因(表S3)。这些基因中的高噪声可能是由描述发生的突发的随机方式引起的([Sanchez and Golding, 2013](#_bookmark41)).

低测序深度时灵敏度和精度饱和

通过进行更深入的测序，可以增强检测低水平转录物的能力。然而，成本和灵敏度之间存在权衡，尤其是对于高通量实验。经验表明，在高通量scRNA-seq实验中，每个细胞可读取10，000–100，000个读数，而对于常规scRNA-seq数据，相应值通常为每细胞~ 100万个读数([Baran-Gale et al., 2018](#_bookmark17)).基于数学-数学模型的先前研究表明，浅层测序(常规深度的1%)也可以提供关于细胞状态的信息([Heimberg et al., 2016](#_bookmark30)).我们对测序数据进行了随机二次抽样，并分析了灵敏度和精密度的相应变化([Figures 5](#_bookmark12)a、5B和S6)。拟合的UMI饱和曲线和基因计数有助于确定适合大多数应用的测序深度。

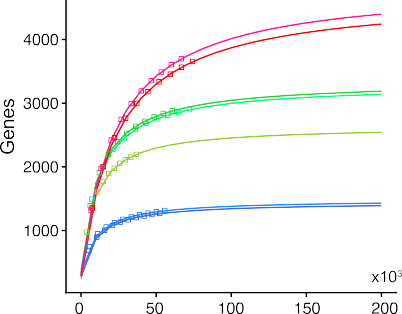
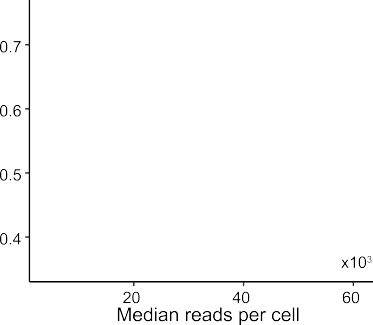
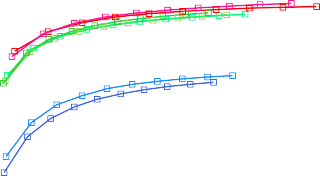
所有的系统在更高的深度都表现出收益递减。对于更灵敏的方法，可以用更少的读数检测相同水平的umi。所有三种方法都可以达到1，1,000 UMIs的阈值，且读取次数少于10K。10X可以检测10，000个umi，其中位值约为20K，但Drop-seq的值为50K。这两项都超过了in- Drop的容量。我们还评估了Drop-seq (~80K)和inDrop(~ 60K；图S6A)。相比之下，由于灵敏度更高，10X需要约200K次读取/单元才能完成。基因数量的检测灵敏度更快饱和。为了达到80%的饱和水平，inDrop或Drop-seq需要~30K读/单元，10X需要~80K读/单元(图S6B)。除了灵敏度之外，精度也决定了系统进行生物发现的分辨率。这里，精度是以一个像元与其他像元之间的最近相关性来衡量的，这也表明了技术噪声的水平。我们研究了测序深度对精确度水平的影响，发现精确度指数随着读取深度的增加(R20，000次有效读取)而迅速饱和

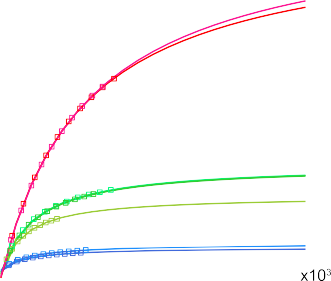
对于所有三个系统([Figure 5](#_bookmark12)c)。这些结果有助于我们建立适当的实证指南——

实验设计的线条。对于最常执行的任务，如单元格类型，20，000次读取/单元格的中位值就足够了。然而，值得注意的是，这些结果来自具有丰富mRNA的细胞系。应根据方案的敏感性和输入的RNA含量考虑所需的测序深度。对于转录活性较低的细胞，如原代细胞，较低的测序深度水平可能足以用于每种方案。

分子细胞73，130-142，2019年1月3日137

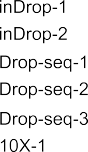
# 甲和乙





# C

基因定量中的偏倚

为了全面比较不同系统描述的转录组，我们使用主成分分析(PCA)和t-分布随机邻域嵌入(tSNE)分析进行降维([Figure 6](#_bookmark13)a)。几乎所有的细胞都根据其起源系统进行了稳健的分离和聚集。尽管同一批次的细胞内存在生物学和技术差异，导致测序读数以及基因和UMI计数存在很大差异，但不同系统之间的差异仍然超过了这些差异的水平。由于以不同的批次和天数处理重复项，批次效应也不明显。在同一系统中，不同批次的数据表现出非常均匀的分布(图S7)。

系统对细胞的分离表明在基因水平上存在系统特异性的定量偏倚。基因水平上mRNA富集的潜在偏差可能与三个主要因素相关:表达丰度(归一化为每百万UMIs基因长度；和GC含量。因此，我们选择了前100个标记基因(见[STAR Methods](#_bookmark50))从每种方法中又分析了这些因素的分布([Fig-ures 6](#_bookmark13)b–6D)。这些基因在生物学复制品之间表现出一致的表达强度。我们发现，与其他系统相比，10X稍微偏向于较短的基因和GC含量较高的基因，而Drop-seq则更容易检测到GC含量较低的基因。这一观察结果与之前的报告相呼应，该报告描述Drop-seq高估了具有低GC比率或长序列的基因的描述([Macoskoet al., 2015](#_bookmark36)).

总之，所有方法在不同批次的技术复制品中表现出非常一致和同质性。这表明合并不同数据的有效性-

图5。通过二次抽样分析得到的不同测序深度时转录组分析灵敏度和噪声水平

(A和B)umi(A)和基因的中位数

* 1. 对于有效读取计数增加的每个样本检测到。
  2. 转录组分析噪声水平随着测序深度快速饱和。噪声测量为最近相关(参见[STARMethods](#_bookmark50)).

另见图S6。

从相同的方法集合在一起。然而，不同方案在基因长度和GC含量方面有明显的偏倚。因此，直接组合这些数据集会引入额外的发散。

讨论

我们使用相同的细胞样本和统一的数据处理管道，比较了三种最广泛使用的基于液滴的高通量单细胞RNA-seq系统inDrop、Drop- seq和10X，以

减少实验设计和数据分析中的偏差。包括技术复制，以确定可能的批次相关工件。对于每个系统，我们对数千个单细胞进行测序。通过对我们的统一数据处理管道中的几个关键参数进行定量分析，我们阐明了每个系统的特点。一般来说，在过滤掉人工因素和错误后，所有三个系统都产生了用于单细胞表达分析的质量数据。细胞分型分析表明批次效应不明显，但与所选系统相关的聚类偏倚明显。这表明，使用来自混合系统的数据集进行细胞分型分析在技术上具有挑战性，应当避免。

在本研究中，我们选择了一种淋巴母细胞样细胞系进行分析，因为细胞系的质量是高度可控的。至少在技术评估方面，我们希望尽可能减少样本质量对所获结果的影响。然而，使用原代细胞，尤其是mRNA含量低的原代细胞进行直接比较，信息会更丰富。为了扩大我们的研究范围，我们进一步用10X系统处理了HEK293细胞，并纳入了三个系统的原始开发者产生的一些数据集([Klein et al.,2015; Macosko et al., 2015; Zheng et al., 2017](#_bookmark33)).如表S5所示，10X显示了更高的灵敏度，从各种细胞中检测到的umi大约是inDrop和Drop-seq do的两倍。inDrop开发人员的结果比我们的要好。我们将这种差异归因于珠质量的批次间差异。如上所示，inDrop细胞条形码的错误率远远高于Drop-seq和10X([Figure 2](#_bookmark9)b)。被贴上有缺陷的条形码会认为这些转录本从一开始就无法检测到。超过一半的inDrop测序数据由于与中的细胞条形码不匹配而被浪费

138分子细胞73，130-142，2019年1月3日

**甲和乙**



**C**





**D**



图6。三个系统中的转录组分析偏倚

(A和B)通过PCA (A)和tSNE (B)聚类的所有三个系统的细胞条形码可视化。

(B–D)转录组分析在基因表达水平(B)、基因长度(C)和GC含量(D)方面存在偏倚。每个系统的前100个标记基因用于演示。还显示了所有基因的分布(灰色)，以供比较。

另见图S7。

我们的数据。其他inDrop用户的反馈显示，不同批次磁珠的等效比例介于25%至65%(未公布数据)。我们还测试了mRNA含量对系统性能的影响。当使用一半HEK293 cDNA用于下游文库制备时，我们检测到大约一半的UMI，与正常HEK293相同(表S5)。所有这些上述结果表明，我们基于淋巴母细胞样细胞系的发现可以推广到其他细胞类型。

对于所有三种系统，磁珠都是由特定制造商专门提供的，很可能很难在小型实验室中生产。因此，珠子的质量，例如它们的大小分散性，对于确定逆转录和进一步反应的稳健性和均匀性特别重要。此外，每个珠子上条形码序列的保真度和纯度也是影响生物信息学管道的关键因素，对于这些因素，应尽量减少伪影和错误。

我们的比较表明，10X通常具有更高的分子灵敏度和精度，以及更低的技术噪声。作为一个更加成熟的商业化系统，本应广泛优化10X方案，这部分反映在磁珠生产的条形码设计和质量控制中。然而，高性能优化也有很高的价格。具体来说，该仪器的成本超过

5万美元，即使不考虑测序成本或仪器折旧，每个细胞的成本也在0.50美元左右(表S6)。

与10X相比，Drop-seq在灵敏度和精度方面稍有折衷，在实验成本方面表现出显著优势，这通常是需要大量单细胞时需要关注的主要问题。作为一个开源系统

(除了珠子)，Drop-seq自2015年推出以来，一直广受欢迎。截至本文撰写之时，Drop-seq协议已被下载近60，000次。构建整个系统不到3万美元。Drop-seq的实验成本约为每个细胞0.10美元(表S6)。因此，鉴于Drop-seq的平衡性能和经济性，它是各个实验室的合理选择。

在某种程度上，inDrop可以被认为是10X的开源版本。两者都使用水凝胶珠进行超泊松加载。它们的珠上引物都是可释放的，以促进转录物的捕获。仪器成本与10X相当，单位电池成本约为10X的一半(表S6)。我们将in- Drop的较低性能归因于其过量的cDNA扩增([Hashimshony et al.,2016](#_bookmark29))以及该协议尚未完全优化的事实。作为一个开源系统，inDrop可采用其他化学方法，并可针对不同类型的RNA-seq方案进行轻松修改。在一项初步实验中，我们在inDrop系统上测试了最广泛使用的scRNA- seq协议Smart-seq2的实现。cDNA图谱与传统的Smart-seq2产品非常相似([Figure 7](#_bookmark14)a)。我们进一步测试了逆转录和cDNA扩增的不同条件。与官方方案产生的结果相似，新数据中的很大一部分(~40%)读数无法分配给真正的细胞条形码。我们经过短暂优化的方案产生的UMI和基因检测结果与官方方案的结果相当([Figures7](#_bookmark14)b–7D；表S4)。尽管转录检测的灵敏度仍然低于其他两种系统，但我们的初步结果证明了inDrop的灵活性，并且该系统对于非标准方法或技术开发是可取的。

分子细胞73，130-142，2019年1月3日139

# A

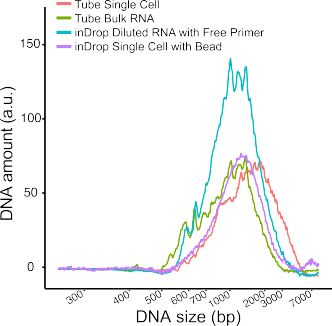
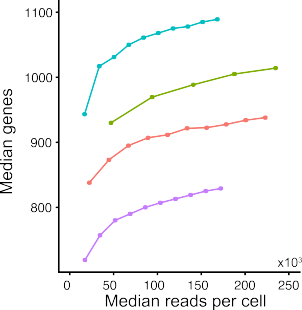
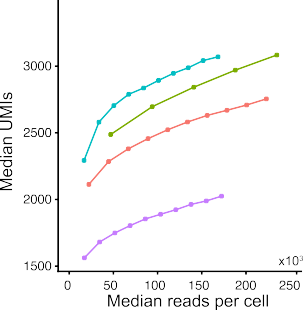
**b、C、D**

图7。在inDrop平台中采用Smart-seq2协议

1. 在试管和inDrop平台中进行的Smart-seq 2 cDNA片段大小比较。(B和C)进行了四种具有不同反应温度和PCR扩增策略的反应(S1、S2、L1和L2，参见[STAR Methods](#_bookmark50)).显示了不同测序深度时的中位检测UMI (B)和基因(C)计数。

(D)在均匀测序深度(100K读数)下四种条件下的UMI分布。L1条件具有较好的敏感性。另见表S4。

鉴于上述所有系统特异性特征，我们提出了指南，以方便选择适合超高通量单细胞研究的基于液滴的scRNA-seq系统。尽管大多数项目使用的细胞数量相对较大，但人类胚胎等珍贵样本需要高效的细胞捕获。细胞盖的超泊松分布对于此类样本可能是必要的。关于转录物检测的实验成本和效率的要求取决于具体情况。一般来说，所有三种系统都提供满意的转录物检测效率，并且更高的效率与更高的实验成本相关。根据经验，10X目前是大多数应用程序的安全选择。当样本丰富时，Drop-seq可能更具成本效益。相比之下，当可选择检测低丰度转录物或需要定制方案时，inDrop成为更好的选择。

STAR+方法

本文在线版提供了详细方法，包括以下内容:

d[KEY RESOURCES TABLE](#_bookmark51)

d[CONTACT FOR REAGENT AND RESOURCE SHARING](#_bookmark52)

d[EXPERIMENTAL MODEL AND SUBJECT DETAILS](#_bookmark53)

b细胞系

d[METHOD DETAILS](#_bookmark54)

b细胞系和细胞制备

b滴序实验

B inDrop实验

B 10X系统实验

b细胞捕获效率的计算

b细胞条形码和UMI分配

b .有效细胞条形码库大小分析

b条形码校正

B有效细胞条形码的测定B基因的对齐和标记读取B UMI校正

b与协议官方分析管道的比较

b基因结构上的阅读分布

b细胞的质量控制和过滤

b有效读取的比例

b .技术噪声分析

b管家基因的技术噪音

b .测序深度二次采样和归一化

inDrop系统中的B- Smart seq 2协议

b .三个系统的公共数据集比较

d[QUANTIFICATION AND STATISTICAL ANALYSIS](#_bookmark55)

b .测序深度拟合和预测

B PCA和tSNE分析

d[DATA AND SOFTWARE AVAILABILITY](#_bookmark56)

补充信息

补充信息包括七个数字和六个表格，可通过本文在线找到，网址为<https://doi.org/10.1016/j.molcel.2018.10.020>。

感谢

我们感谢吴梵测试数据处理管道。这项工作得到了中国科技部(2016YFC0900100 to J.W .和2018YFA0108100 to Y.H .)、中国国家自然科学基金(21675098 to J.W .和21327808和21525521 to Y.H .)以及北京基因组高级创新中心的支持。

作者贡献

Y.H .和J.W .构思了这个项目。T.L .、F.L .和J.Y .进行了实验。X.Z .、Z.L .、Y.C .和J.Y .分析的数据。所有作者都参与了手稿的准备工作。

利益声明

作者宣布没有相互竞争的利益。

收到:2018年5月2日

修订:2018年8月25日

接受:2018年10月12日

发布日期:2018年11月21日

140分子细胞73，130-142，2019年1月3日

参考

[Abate，A.R .、Chen、C.H .、Agresti、J.J .和Weitz，D.A. (2009)。击败泊松使用密集排序的封装统计。实验室芯片9，2628-2631](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref1)。

[Agresti，J.J .、Antipov、e .、A.R .、Ahn、k .、Rowat、A.C .、Baret、J.C .Marquez，m .，Klibanov，A.M .，Griffiths，A.D .，和Weitz，D.A. (2010)。超高-用于定向进化的基于液滴的微流体中的通量筛选。继续。纳特。阿卡德。Sci。美国107，4004–4009](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref2)。

[Baran-Gale，j .，Chandra，t .，和Kirschner，K. (2018)。的实验设计单细胞RNA测序。简短。Funct。基因组学17，233-239](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref3)。

[Brennecke，p .，Anders，s .，Kim，J.K .，Ko1odziejczyk，A.A .，Zhang，x .，Proserpio，动词 （verb的缩写），Baying，b .，Benes，v .，Teichmann，S.A .，Marioni，J.C .，和Heisler，M.G .(2013).单细胞RNA-seq实验中技术噪声的说明。纳特。方法10，1093–1095](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref4)。

[Briggs，J.A .、Weinreb，c .、Wagner、D.E .、Megason、s .、Peshkin、l .、Kirschner、M.W .，和Klein，A.M. (2018年)。脊椎动物基因表达的动力学单细胞分辨率的胚胎发生。科学360，eaar5780](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref5)。

[Buettner，f .，Natarajan，K.N .，卡萨勒，F.P .，Proserpio，v .，Scialdone，a .，泰斯，F.J .、Teichmann、S.A .、Marioni、J.C .和Stegle、O. (2015年)。计算的单细胞RNA测序数据中细胞间异质性的分析揭示隐藏的细胞亚群。纳特。Biotechnol。33, 155–160](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref6)。

[Butler，a .，Hoffman，p .，Smibert，p .，Papalexi，e .，和Satija，R. (2018)。(使)融入跨不同条件、技术和的单细胞转录组学数据物种。纳特。Biotechnol。36, 411–420](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref7)。

[Deng，q .，Ramsko ld，d .，Reinius，b .，Sandberg，R. (2014)。单细胞RNA-seq揭示哺乳动物中动态随机单等位基因表达细胞。科学343，193-196](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref8)。

[Dixit，a .，Parnas，o .，Li，b .，Chen，j .，Fulco，C.P .，Jerby-Arnon，l .，N.D .，Dionne，d .，Burks，t .，Raychowdhury，r .，et al. (2016)。扰动-序列:使用池化的可扩展单细胞RNA分析剖析分子电路基因筛选。167号牢房，1853–1866 . e17](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref9)。

[Dobin，a .，Davis，C.A .，Schlesinger，f .，Drenkow，j .，Zaleski，c .，Jha，s .，Batut，页（page的缩写），Chaisson，m .，和Gingeras，T.R. (2013)。STAR:超快通用RNA-seq校准器。生物信息学29，15-21](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref10)。

[Duncombe，T.A .、Tentori，A.M .、Herr，A.E. (2015年)。微流体:重构生物调查。纳特。Mol牧师。Cell Biol。16, 554–567](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref11)。

[范汉忠，傅，郭国凯，，S.P. (2015)。表达式分析。组合的用于基因表达细胞术的单细胞标记。科学347，1258367](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref12)。

[Farrell，J.A .、Wang，y .、Riesenfeld、S.J .、Shekhar、k .、Regev、a .和Schier、](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref13)

[A.F. (2018年)。发育轨迹的单细胞重建斑马鱼胚胎发生。科学360，eaar3131](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref13)。

[Gr](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref14)u[n，d .，和van Oudenaarden，A. (2015年)。单细胞的设计与分析测序实验。163，799–810室](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref14)。

[Han，x .，Wang，r .，Zhou，y .，Fei，l .，Sun，h .，Lai，s .，Saadatpour，a .，Zhou，z .陈，h .，叶，f .，等(2018)。通过Microwell-seq绘制小鼠细胞图谱。172，1091–1107 . e17室](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref15)。

[Hashimshony，t .，Wagner，f .，Sher，n .，Yanai，I. (2012)。CEL-seq: single-通过多重线性扩增的细胞RNA-seq。电话:2，666-673](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref16)。

[Hashimshony，t .、Senderovich，n .、Avital，g .、Klochendler，a .、德·勒乌，y .Anavy，l .，Gennert，d .，Li，s .，Livak，K.J .，Rozenblatt-Rosen，o .，et al. (2016)。CEL-Seq2:高度敏感的多重单细胞RNA-Seq。Genome Biol。*17, 77*](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref17)。

[Heimberg，g .，Bhatnagar，r .，El-Samad，h .，和Thomson，M. (2016)。低的基因表达数据的维数使得能够准确提取来自浅层测序的转录程序。Cell Syst。2, 239–250](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref18)。

[Jaitin，D.A .、Kenigsberg、e .、Keren-Shaul、h .、Elefant、n .、Paul、f .、Zaretsky、I、Mildner，a .，Cohen，n .，Jung，s .，Tanay，a .，Amit，I. (2014)。非常平价用于将组织无标记分解成细胞等位单细胞RNA-seq类型。科学343，776–779](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref19)。

[Kivioja，t .，V](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref20)a[h](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref20)a[rautio，a .，Karlsson，k .，Bonke，m .，Enge，m .，Linnarsson，s .和Taipale，J. (2011)。利用唯一性计算分子的绝对数量分子标识符。纳特。方法9，72-74](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref20)。

[Klein，A.M .、Mazutis，l .、Akartuna，I .、Tallapragada，n .、Veres，a .、Li，v .Peshkin，l .，Weitz，D.A .，和Kirschner，M.W. (2015)。sin-的液滴条形码应用于胚胎干细胞的单细胞转录组学。161，1187–1201室](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref21)。

[Lake，B.B .，Ai，r .，Kaeser，G.E .，Salathia，N.S .，Yung，Y.C .，Liu，r .，Wildberg，a .高，d，冯，H.L .，陈，s .，等(2016)。神经元亚型和多样性由人脑的单核RNA测序揭示。科学352，1586–1590](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref22)。

[Li，h .，Durbin，R. (2009)。快速准确的短读对齐布伦斯-惠勒转型。生物信息学25，1754–1760](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref23)。

[Macosko，E.Z .，Basu，a .，Satija，r .，Nemesh，j .，Shekhar，k .，Goldman，m .Tirosh，I .，Bialas，A.R .，Kamitaki，n .，Martersteck，E.M .，et al. (2015)。高度平价使用纳升滴-的单个细胞的等位基因全基因组表达谱分析让我们。161，1202–1214室](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref24)。

[Olsson，a .，Venkatasubramanian，m .，Chaudhri，V.K .，Aronow，B.J .，Salomonis，n .，Singh，h .，和Grimes，H.L. (2016年)。的单细胞分析导致二元细胞命运选择的混合谱系状态。Nature 537，698-702](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref25)。

[Papalexi，e .，和Satija，R. (2018)。单细胞RNA测序探索im-mune细胞异质性。纳特。Rev. Immunol。18, 35–45](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref26)。

[Patel，A.P .、Tirosh，I .、特隆贝塔、J.J .、Shalek、A.K .、Gillespie、S.M .、Wakimoto、H.，Cahill，D.P .，Nahed，B.V .，Curry，W.T .，Martuza，R.L .，et al. (2014)。单一的细胞RNA-seq强调原发性胶质母细胞瘤的肿瘤内异质性。科学344，1396–1401](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref27)。

[Picelli，s .，Faridani，O.R .，Bjo rklund，A.K .，Winberg，g .，Sagasser，s .，和r . sand Berg(2014)。使用Smart-从单细胞中获得的全长RNA-seqseq2。纳特。协议。9, 171–181](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref28)。

[花粉，A.A .，Nowakowski，T.J .，Shuga，j .，Wang，x .，Leyrat，A.A .，Lui，J.H .，Li，名词（noun的缩写），Szpankowski，l .，Fowler，b .，Chen，p .，et al. (2014)。低承保单-细胞mRNA测序揭示细胞异质性和激活信号大脑皮层发育的途径。纳特。Biotechnol。32, 1053–1058](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref29)。

[Prakadan，S.M .，Shalek，A.K .，和Weitz，D.A. (2017)。通过收缩进行缩放:em-用微流体装置为单细胞“组学”提供动力。纳特。热内特牧师。18,345–361](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref30)。

[Ramsko ld，d .，lo，s .，Wang，Y.C .，Li，r .，Deng，q .，Faridani，O.R .，Daniels，G.A .，Khrebtukova，I .，Loring，J.F .，Laurent，L.C .，et al. (2012)。全长来自RNA单细胞水平和单个循环肿瘤细胞的mRNA-seq。纳特。Biotechnol。30, 777–782](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref31)。

[桑切斯和戈尔丁(2013)。遗传决定因素和细胞因子嘈杂基因表达的束缚。科学342，1188–1193](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref32)。

[Semrau，s .，Goldmann，J.E .，Soumillon，m .，Mikkelsen，T.S .，Jaenisch，r .，和van Oudenaarden，A. (2017年)。揭示的谱系承诺动态分化胚胎干细胞的单细胞转录组学。纳特。通信。8, 1096](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref33)。

[Shalek，A.K .、Satija、r .、Shuga、j .、J.J .、Gennert、d .、Lu、d .、Chen、页（page的缩写），Gertner，R.S .，Gaublomme，J.T .，Yosef，n .，et al. (2014)。单细胞RNA-seq揭示了细胞变异的动态旁分泌控制。自然510，363–369](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref34)。

[街道，A.M .，黄，Y. (2014)。单细胞RNA的深度有多深序列？纳特。Biotechnol。32, 1005–1006](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref35)。

[街道，A.M .、张、x .、曹、c .、庞、y .、吴、x .、熊、l .、杨、l .、付、y、赵，l .，唐，f .，黄，Y. (2014)。微流控单细胞全反脚本体测序。继续。纳特。阿卡德。Sci。美国111，7048–7053](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref36)。

[Svensson，v .，Natarajan，K.N .，Ly，L.H .，Miragaia，R.J .，Labalette，c .Macaulay，I.C .，Cvejic，a .，和Teichmann，S.A. (2017)。罪恶的力量分析-单细胞RNA测序实验。纳特。方法14，381–387](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref37)。

[Svensson，v .，Vento-Tormo，r .，和Teichmann，S.A. (2018年)。指数的过去十年中单细胞RNA-seq的缩放比例。纳特。协议。13, 599–604](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref38)。

[Tanay，a .，和Regev，A. (2017)。从现象扩展单细胞基因组学-机械技术。Nature 541，331–338](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref39)。

[Tang，f .，Barbacioru，c .，Wang，y .，Nordman，e .，Lee，c .，Xu，n .，Wang，x .Bodeau，j .，Tuch，B.B .，Siddiqui，a .，等人(2009年)。mRNA-seq全转录组-单细胞的时间分析。纳特。方法6，377–382](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref40)。

分子细胞73，130-142，2019年1月3日141

[Tirosh，I .，Venteicher，A.S .，Hebert，c .，Escalante，L.E .，Patel，A.P .，Yizhak，k .Fisher，J.M .，Rodman，c .，Mount，c .，Filbin，M.G .，et al. (2016)。单细胞RNA-seq支持人少突胶质细胞瘤的发育分级。Nature 539，309–313](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref41)。

[Treutlein，b .、Brownfield、D.G .、Wu、A.R .、Neff、N.F .、Mantalas、G.L .、Espinoza、F.H .，Desai，T.J .，Krasnow，M.A .，和Quake，S.R. (2014)。重建线条-使用单细胞RNA-seq的远端肺上皮的年龄分级。自然*509, 371–375*](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref42)。

[Venteicher，A.S .、Tirosh，I .、Hebert，c .、Yizhak，k .、Neftel，c .、Filbin、M.G .Hovestadt，v .，Escalante，L.E .，Shaw，M.L .，Rodman，c .，等人(2017年)。IDH突变胶质瘤的解耦遗传学、谱系和微环境通过单细胞RNA-seq。科学355，eaai8478](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref43)。

[Villani，A.C .、Satija、r .、Reynolds、g .、Sarkizova、s .、Shekhar、k .、Fletcher、j .Griesbeck，m .、Butler，a .、郑，s .、Lazo，s .、等(2017)。单细胞RNA-seq揭示了新类型的人血树突细胞、单核细胞和前天才。科学356，eaah4573](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref44)。

[Wagner，D.E .、Weinreb，c .、Collins、Z.M .、Briggs、J.A .、Megason、S.G .、和Klein，A.M. (2018年)。基因表达图谱和斑马鱼胚胎的谱系。科学360，981–987](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref45)。

[Wu，A.R .、Neff、N.F .、Kalisky、t .、Dalerba、p .、Treutlein、b .、Rothenberg、M.E .](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref46)

[Mburu，F.M .、Mantalas、G.L .、Sim、s .、Clarke、M.F .、Quake、S.R. (2014)。单细胞RNA测序方法的定量评估。纳特。方法11、41-46](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref46)。

[吴，A.R .，王，j .，Streets a . m .，Huang，Y. (2017)。单细胞转录组-国家分析。安妮。肛门牧师。化学。(加利福尼亚州帕洛阿尔托)10，439-462](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref47)。

[阎，l，杨，m，郭，h，杨，l，吴，j，李，r，刘，p，连，y，郑，x，严，j .，等(2013)。人着床前单细胞RNA-seq分析胚胎和胚胎干细胞。纳特。结构。摩尔。Biol。20, 1131–1139](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref48)。

[Zheng，G.X .、Terry、J.M .、bel flanger、p .、Ryvkin、p .、Bent、Z.W .、Wilson、r .](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref49)

[Ziraldo，S.B .，Wheeler，T.D .，McDermott，G.P .，Zhu，j .，et al. (2017)。单细胞的大规模并行数字转录分析。纳特。通信。8, 14049](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref49)。

[Ziegenhain，c .，Vieth，b .，Parekh，s .，Reinius，b .，Guillaumet-Adkins，a .，Smets，米（meter的缩写）），莱恩哈特，h .，Heyn，h .，Hellmann，I .，和Enard，W. (2017)。比较的单细胞RNA测序方法分析。摩尔。65，631–643 . E4室](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref50)。

[Zilionis，r .，Nainys，j .，Veres，a .，Savova，v .，Zemmour，d .，Klein，A.M .，和l . mazutis(2017年)。使用液滴显微技术进行单细胞条形码和测序射流学。纳特。协议。12, 44–73](http://refhub.elsevier.com/S1097-2765(18)30880-3/sref51)。

142分子细胞73，130-142，2019年1月3日

STAR+方法

关键资源表

试剂或资源来源标识符

关键商业分析

|  |  |  |
| --- | --- | --- |
| 滴液珠 | Chemgenes Barcoded Bead SeqB | Cat#MACOSKO-2011-10 |
| INDROP单细胞RNA SEQ试剂盒 | 1cellbio，Inc | 不适用的 |
| 铬单细胞30文库和凝胶珠试剂盒v2 | 10X基因组学 | 类别编号PN-120237 |

存储的数据

|  |  |  |
| --- | --- | --- |
| 来自Drop-seq原纸的原始数据 | NCBI地球观测组织 | GEO: GSE63472 |
| inDrop原纸的原始数据 | NCBI地球观测组织 | GEO: GSE65525。 |
| 原始数据来自10倍的原纸 | 10基因组学 | <https://support.10xgenomics.com/single-cell-gene-expression/datasets> |
| 原始和处理的数据集 | 本文 | GEO: GSE111912 |

实验模型:细胞系软件和算法

GM12891细胞系Coriell Institute N/A

|  |  |  |
| --- | --- | --- |
| 稀有 | r核心团队 | [http://www.R-project.org](http://www.R-project.org/) |
| 修拉 | Rahul Satija | <https://satijalab.org/seurat/> |
| 星星 | 亚历山大·多宾 | <https://github.com/alexdobin/STAR> |
| baseqDrops | 本文 | <https://github.com/beiseq/baseqDrops> |
| Samtools | [Li and Durbin, 2009](#_bookmark34) | <http://samtools.sourceforge.net/> |
| 细胞游侠 | 10基因组学 | <https://github.com/10XGenomics/cellranger> |

试剂和资源共享联系人

有关资源和试剂的更多信息和请求应直接发送给王建斌的负责人联系人并由其完成([jianbinwang@tsinghua.edu.cn](mailto:jianbinwang@tsinghua.edu.cn)).

实验模型和受试者详情

细胞系

GM12891细胞系(雄性)购自Coriell Institute。使用含有L-谷氨酰胺(cat)的RPMI-1640培养基制备完全生长培养基。编号11875-085；Life Technologies)，10%胎牛血清(猫)。编号16000-044；Life Technologies)，以及1%青霉素和链霉素。将细胞系与5%二氧化碳在37℃的培养瓶中培养。

方法详情

细胞系和细胞制备GM12891细胞系购自Coriell Institute。使用含有L-谷氨酰胺(cat)的RPMI-1640培养基制备完全生长培养基。编号11875-085；Life Technologies)，10%胎牛血清(猫)。编号16000-044；Life Technologies)，以及1%青霉素和链霉素。将细胞系与5%二氧化碳在37℃的培养瓶中培养。细胞浓度维持在5×105至1×106个细胞/mL之间。实验前，显微镜下确认细胞大体情况。规则的圆形细胞和一些细胞聚集体表明细胞状态良好。以150 g离心5 min收集细胞，然后用血细胞计数器计数。

拖放实验

用PBS-BSA洗涤细胞三次，并用40mm细胞滤水器过滤。然后计数细胞并将浓度调节至100个细胞/mL。所有后续步骤均按Macosko等人的在线详细说明执行。(<http://mccarrolllab.org/dropseq/>).简而言之，我们在PBS-BSA中加载了造滴油、细胞和条形码珠(cat。编号:MACOSKO-2011-10；Chemgenes Barcoded Bead SeqB)在裂解缓冲液中滴入液滴，生成微流控装置。细胞在液滴中裂解，释放mRNA。磁珠捕获液滴中的mRNAs。破乳后，将珠子混合在一起。我们进行了逆转

分子细胞73，130–142 . e1–E5，2019年1月3日E1

转录和ExoI消化。将2000个珠子(或100个STAMPs)等分到一个PCR管中进行PCR扩增。将PCR产物混合在一起，并使用AMPure XP磁珠进行纯化。我们进一步构建了文库，并使用定制的Drop-seq read 1引物对Illumina HiSeq 4000进行测序。

inDrop实验

我们从1CellBio购买了inDrop仪器和水凝胶珠。我们按照制造商的方案执行所有步骤:硅烷化方案v4.0(<https://1cell-bio.com/wp-content/uploads/2017/10/Silanization-Protocol-v4.pdf>)，inDrop单细胞逆转录协议v2.1(<https://1cell-bio.com/wp-content/uploads/2017/11/inDrop-Single-Cell-Encapsulation-and-Reverse-Transcription-Protocol-Version-2.1.pdf>)，以及inDrop库制备方案v1.2(<https://1cell-bio.com/wp-content/uploads/2017/03/InDrop_LibraryPrep_Protocol_v1.2.pdf>).简而言之，在每次实验之前，我们对微流控芯片进行了硅烷化，并对水凝胶珠进行了预处理。将制滴油、细胞再悬浮和RT/裂解缓冲液加载到芯片中以生成液滴。将乳液收集在冰上的试管中，并通过紫外线照射以释放底漆。反转录在液滴中进行。破乳后，过滤水凝胶珠。RT产物通过ExoI/HinfI消化，并使用AMPure XP磁珠纯化。第二链cDNA采用NEB第二链合成试剂盒合成。IVT后，RNA产物被片段化，并通过随机引物进行逆转录。用AMPure XP磁珠纯化产物，并用qPCR进行定量。我们进一步构建了文库，并在Illumina HiSeq 4000上进行了测序。

10X系统实验

我们按照10X方案执行了所有步骤。我们使用了铬单细胞30文库和凝胶珠试剂盒v2 (10X基因组学)。简而言之，所有样本和试剂都已制备好并装载到芯片中。然后，我们运行铬控制器进行液滴生成。在液滴中进行逆转录。通过破乳和磁珠纯化回收cDNA。将预扩增的cDNA进一步进行文库制备。在Illumina Hiseq 4000上对文库进行测序。

细胞捕获效率的计算如上所述计数加载到每个系统中的细胞数。使用通过质量阈值的细胞条形码数(总UMIs R，000，最近相关R 0.6)测定捕获的细胞数。

细胞条形码和UMI分配

对于10X，我们通过从读取1序列中提取前16 bp和后8 bp，从每个读取对中获得细胞条形码和UMI。同样，可以访问Drop-seq的条形码。inDrop的条形码设计更为复杂，因为全细胞条形码包含两个部分(命名为CB1和CB2)，由一个名为W1的22 BP间隔序列分隔。CB1的长度范围为8至12 bp。我们首先通过容忍两个错配来定位W1序列。然后，我们可以确定CB1的长度和整个细胞的条形码序列。我们汇总了检索到的细胞条形码，并导出了计数数据以供下游分析。对于10X和inDrop，通过要求条形码出现在相关协议的条形码白名单中，进一步过滤了条形码。

有效细胞条形码库大小分析

通过比较和计数多个实验数据集之间的常见条形码，估计有效条形码库。通过使用泊松分布模型，理论上可以通过单个实验来估计条形码库的大小。然而，它需要精确数量的输入单元，这实际上是不可能获得的。因此，我们采用了成对分析。我们假设有效或真实的条形码库是所设计的整个库的子集，并且在不同实验中是一致的。在每份样本中，从有效库中随机抽取检测到的细胞条形码(无需替换)；碰巧的是，应该有一些条形码在一个以上的样本中重复出现，这被称为条形码冲突。具有给定检测细胞数的两个样本之间的碰撞数主要由实际条形码库大小决定并反映。

通过多次(10000次)运行模拟，估计了在要求的库大小下样本之间的理想碰撞数。对于每次模拟，我们分别根据两个样本中的细胞编号，从所要求的库大小的条形码池中采样(不替换)给定数量的细胞条形码。所有模拟中两个样本之间重复出现的条形码的平均数被用作估计的碰撞数。对于inDrop，估计碰撞次数为28次，但观察显示为92次。对于包含三个样本的Drop-seq，Drop-seq-3和Drop-seq-2的条形码将合并。观测值为零，而模拟报告的平均值为0.3。对于10X，通过模拟，从734，000个细胞条形码库中采样1560和6478个细胞条形码时，平均碰撞数是13.7。观察到的碰撞数是22。

我们进一步计算了在一系列有效库大小下检测到这种数量的冲突的可能性(范围从要求大小的5%到100%，间隔5%)。对于每个假定的文库大小，在这种文库大小下的两个实验中，使用细胞条形码的数量进行了10，000次随机取样。所有模拟结果的观测值中报告与我们实验中相同数量的碰撞的比例被称为相应库大小的可能性。基于我们的实验观察，李可利-胡德分析有助于指出实际图书馆规模的可能性。

e2 Molecular Cell 73，130–142 . E1–E5，2019年1月3日

条形码校正

理想情况下，同一磁珠上的细胞条形码序列应相同。然而，由于DNA合成中的错误，细胞条形码序列中存在错配甚至缺失。这些错误将导致同一细胞的测序读数被拼接到其他条形码中，从而扩大细胞数量。我们采用一种相对简单的方法来纠正各种错误。具体而言，按丰度对原始条形码进行分类，并将1-bp不匹配范围内的细胞条形码聚合成具有较高读取计数的共有条形码。聚合后，每个细胞条形码由原始读数和更正(但包含条形码错误)读数组成。计算校正读数相对于每个条形码总读数的比例，并定义为细胞条形码错误率。统计分析采用读取20K以上的细胞条码的错误率。

确定有效的细胞条形码

校正后的单元格条码数量大大超过了输入的单元格数量，称为“条码膨胀”。这些细胞条形码大多来源于含有小珠但没有细胞的液滴。我们假设来自高质量细胞的有效细胞条形码应具有足够的mRNA分子和更高的读取计数。有多种方法可以通过读取的计数数据推断出有效的细胞条形码。我们采用多种策略来确定细胞数量(图S3)。我们首先通过考虑细胞输入、细胞捕获率和其他因素(称为“估计”)，获得捕获细胞数量的粗略估计。还确定了读取次数超过20K和40K的细胞数量(称为读取R 20/40K)。我们还遵循了10X流水线(Cell Ranger)中使用的方法，该方法为细胞的umi设置了一个阈值。假设前1%的细胞含有的umi约是典型细胞的10倍。得到的估计细胞数称为UMI (1/10)。我们还在经过对数转换的条形码等级和条形码读数图中观察到一个明显的拐点([Figure 2](#_bookmark9)d)。我们生成了10倍样本的经对数转换的条形码等级和条形码读数图(图S3A)，并在估计的细胞数量周围观察到一个陡峭的悬崖。然而，这种现象在inDrop和Drop-seq样本中分别更为微妙或不存在。

以上讨论的方法都基于某些假设或假定，这些假设或假定可能不适用于所有方法和细胞组成。这里，我们通过质量控制确定每个样本中的实际细胞数，并基于实验信息通过粗略估计对细胞数进行过滤。要求有效细胞至少有1，000个umi，最接近的相关关系应超过0.6(参见下面的方法)。

基因的比对和标签阅读

使用STAR(RNA-seq数据的高效对齐器)将读数与参考基因组(GRCh38)对齐([Dobin et al.,2013](#_bookmark23)).大多数读取可以唯一映射，并且可以根据注释对读取进行标记。要求读数至少有50%的长度被绘制出来并与外显子区域重叠。对于多个对齐的读数，读数应来自要计数的同一基因。

UMI更正

umi也可能受到测序错误的影响。在我们的过程中，来自同一细胞条形码的同一基因中的umi按其计数进行分类。将1汉明距离内的umi进行聚合。我们观察到，对于某些具有成百上千UMI的高表达基因，UMI校正所需的时间可能呈指数级增加。默认情况下，我们禁用了那些UMI计数大于100的基因的关联。

与协议官方分析管道的比较

官方分析管道分别在他们的指导下下载和执行。我们组合了UMI表，为官方管道和我们的多功能管道提供了大约100个细胞条形码(图S2A)。比较两种分析的基因表达水平。我们还计算并比较了两个管道中每个细胞条形码的umi总数(图S2B)。

Picard工具对基因结构的读取分布(<http://broadinstitute.github.io/picard/>)及其RNASeqMetrics命令用于分析基因组和基因体上的map- ped读数的分布。GRCh38的基因结构注释以refFlat格式从UCSC基因组浏览器下载。对以下区域进行了计数:编码区、UTRs区、内含子区和基因间区。

细胞的质量控制和过滤

UMI计数反映了转录组的分子多样性。我们评估了所有细胞的umi基本含量和数量。合格的细胞条形码应包含1，000个以上umi。我们进一步使用前500个基因的UMI计数生成了所有有效细胞条形码的Spearman相关矩阵。然后，我们获得每个细胞条形码的最近邻，并计算最大成对相关性。这些值显示异常值，因此移除了相关性最近< 0.6的细胞条形码。

分子细胞73，130–142 . E1–E5，2019年1月3日e3

有效读取的比例过滤前，从所有估计的条形码中读取的估计条形码读取计数，映射到基因组的读取称为可映射读取。UMI效应读取计数映射到有助于分子计数的基因体区域的读取。其余读数通过两次质量控制过滤进一步过滤。

技术噪声分析通过计算所有有效成对条形码的Spearman相关系数进行最近相关分析。为了在基因水平上评估技术噪声，我们从每个样本中随机选择了500个细胞及其前1000个高表达基因。为了调整分子数和捕获效率的差异，通过乘以一个因子使UMI归一化，该因子使得每个细胞的归一化UMI计数之和等于1，000，000(umi/百万)。

看家基因的技术噪音看家基因(HK基因)列表可从以下网站下载[https://www.tau.ac.il/](https://www.tau.ac.il/%7Eelieis/HKG/)~[elieis/HKG/](https://www.tau.ac.il/%7Eelieis/HKG/)。持家基因和非持家基因(非HK基因)用不同的颜色标记，以描述它们在总体噪声水平方面的分布。HK基因与最低量的生物噪声相关，其总体噪声水平可近似视为技术噪声。香港与非香港基因分布的相似性表明，生物噪声水平显著低于技术噪声水平。

时序深度二次采样和归一化

为了确保公平比较三种方案的性能，而不考虑它们的不同测序深度，我们进行了二次抽样分析。我们对原始测序读数进行了二次采样，以获得10%至90%的原始测序深度，采样间隔为5%。所得数据随后经过与上述相同的处理流水线。我们获得了相应的细胞条形码、中位原始测序深度、umi和每个测序深度的基因。由于Drop-Seq-1的测序深度相对较低(~37K读数)，因此我们选择了每个样本的二次抽样比率，以使中位测序深度(原始读数)接近37，000。

inDrop系统中的Smart-seq2协议

我们在inDrop平台上测试了基于Smart-seq2的协议。在液滴生成步骤中，将上标III替换为上标II (10 mL/100 mL RT进样口)，并向100 mL RT进样口中加入2mL 100mM模板Switch Oligo (TSO，AAGCAGTGGTATCAACGCAGAGTA CATrGrGrG)。在逆转录步骤，测试了两种策略。一项涉及暴露于42℃1.5小时(42℃RT)，另一项涉及42℃1.5小时，然后是50℃2分钟和42℃2分钟，共10个循环(42/50℃RT)。逆转录后，按照标准inDrop方案的相同步骤进行破乳。水相用0.6±3安培的磁珠纯化。在cDNA扩增步骤中，测试了两种策略。一项涉及将所有cDNA汇集在一起进行扩增(Pool PCR)，另一项涉及将cDNA分成每管约200个回收细胞的试管，并在扩增后汇集在一起(Split PCR)。

三个系统公共数据集的比较我们从最初发表的论文中优先收集了这三个系统的多个数据集。对于没有原始测序读取文件的前瞻性数据集，我们使用了从Gene Expression Omnibus (GEO)下载的经过处理的UMI表达表。对原始测序文件进行二次取样，以获得相对均匀的测序深度(每个细胞约30K读数)。我们还纳入了在本实验室制备的10X HEK 293样本。破乳后的cDNAs被分成两半，用于下游反应。因此，对于每一个条形码，其所含的量约为正常HEK293细胞的一半。我们获得了下列数据集的原始文件(fastk/SRA):inDrop的小鼠es细胞([Klein et al., 2015](#_bookmark33))，脑细胞减少10倍，pan T细胞减少10倍，HEK293减少-seq([Macosko et al., 2015](#_bookmark36))和10X。使用我们自己开发的管道对这些数据集进行了重新分析，以获得数字表达矩阵。为了识别有效的细胞条形码，使用UMIs > 1000的标准过滤所有数据集。不可用的值用“NA”标记。''

相同类型的细胞在不同的系统之间进行了更公平的比较。所有系统都存在相似阶段的斑马鱼数据集。inDrop和10X可通过pan T细胞或T调节细胞进行比较。Drop-seq和10X可使用其HEK293数据集进行比较。

定量和统计分析

测序深度拟合和预测

当在不同深度对样本进行测序时，会干扰样本和方法之间的比较。因此，我们将原始读数随机二次采样至不同的测序深度(10%至90%，间隔10%)，并测量其性能。如上所述计算检测到的umi和基因的数量以及每个深度的精确度水平指数。使用以下方程进行基因和umi饱和水平的拟合和预测。

e4 Molecular Cell 73，130–142 . E1–E5，2019年1月3日

*y = a +*

*b x + c*

PCA和tSNE分析

我们使用的是Seruat包装(<https://satijalab.org/seurat/>) ([Butler et al., 2018](#_bookmark21))用于PCA和tSNE分析。为了提高效率，我们从每个样本中随机抽取了500个细胞。该管道通过细胞和基因过滤、数据标准化和最易变基因的识别进行。该管道还会报告每个样本的标记基因，这些基因的表达量高于其他样本。我们为每个样本生成了前100个标记基因，这些基因根据它们的系统进行了聚合。每个基因的长度和GC含量是通过对所有相应转录物的值取平均值来计算的。

数据和软件可用性

本文中报告的数据的注册号为GEO: GSE111912。数据处理管道(baseqDrops)可在github(<https://github.com/beiseq/baseqDrops>).

分子细胞73，130–142 . E1–e5，2019年1月3日E5

分子细胞，第73卷

补充信息

液滴基复合材料的对比分析

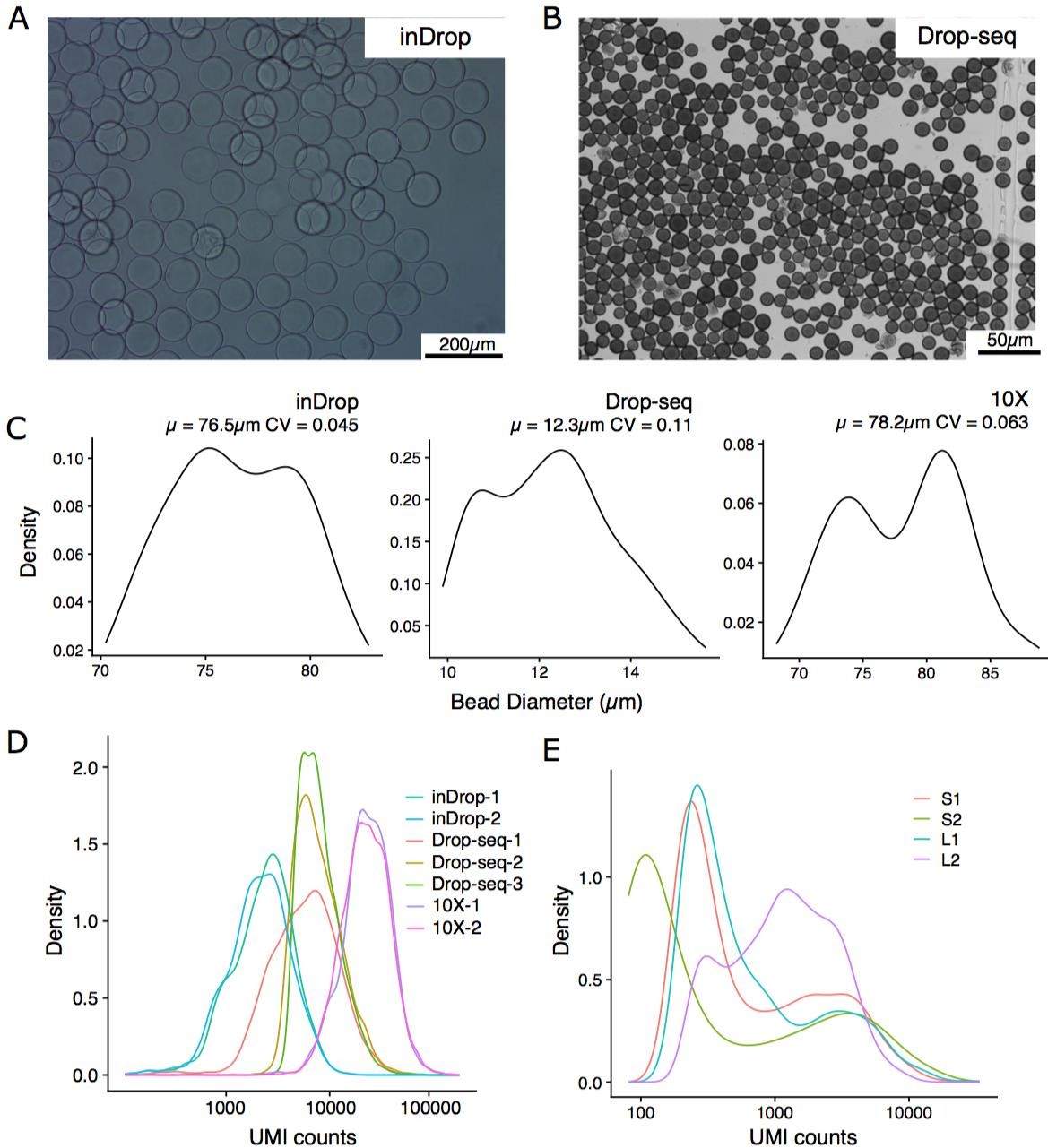
超高通量单细胞RNA-Seq系统

张先念、李天棋、、姚家成、李泽耀、黄、等

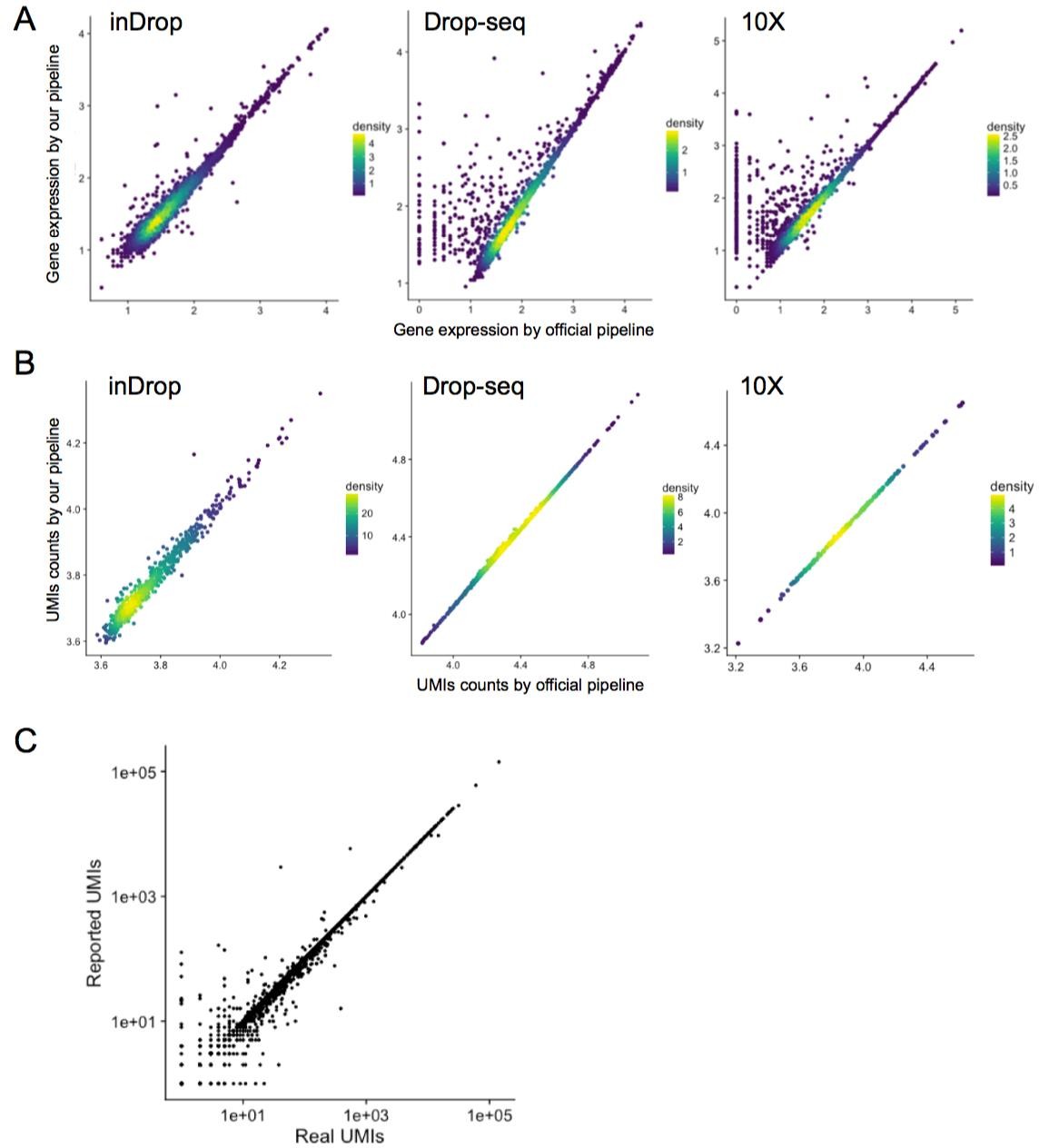
**补充数字和表格**

基于液滴的超高通量单细胞RNA-seq系统的比较分析

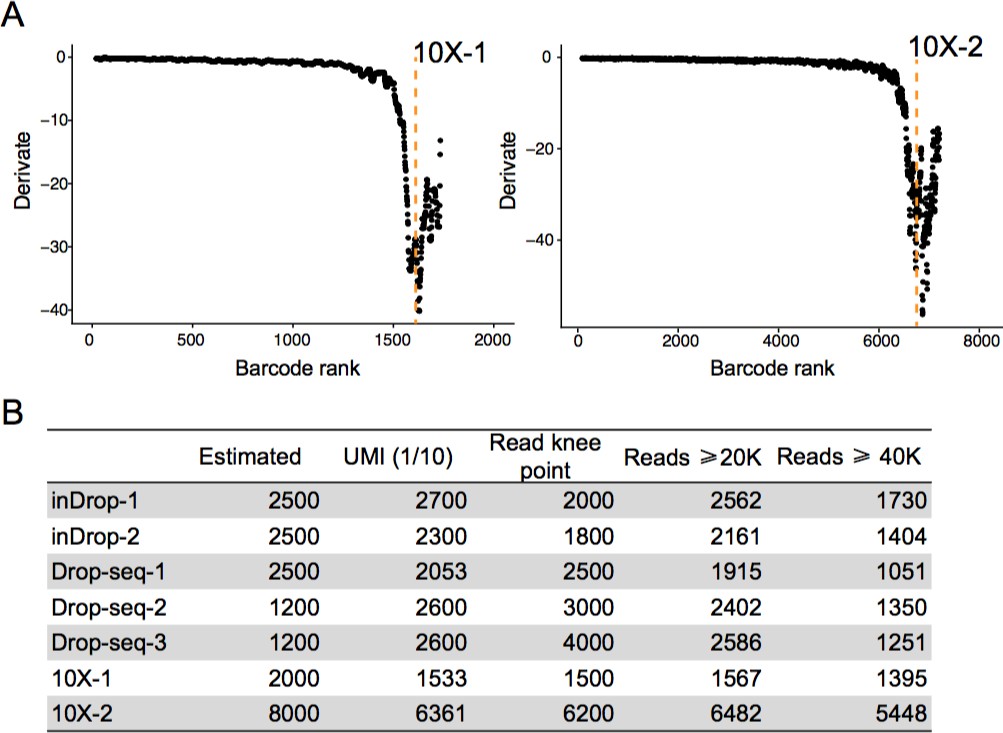
张先念，1，#李天棋，2，，3，，，4，#姚家成，2李泽耀，5黄，1 \*，王建斌，



图S1(与图1和图7相关)。Drop-seq和inDrop磁珠的大小分布。不同放大倍率下Drop-seq磁珠(A)和inDrop磁珠(B)的明视野显微镜图像。(C)通过图像处理测量磁珠直径，并绘制inDrop、Drop-seq和10X的磁珠直径密度分布图。注明了胎圈直径的平均值和变异系数(CV)。(D)所有样本的UMI分布均遵循单峰分布。(E)四种Smart-seq2 inDrop样本的UMI分布。图7显示了样本名称及其对应的条件。

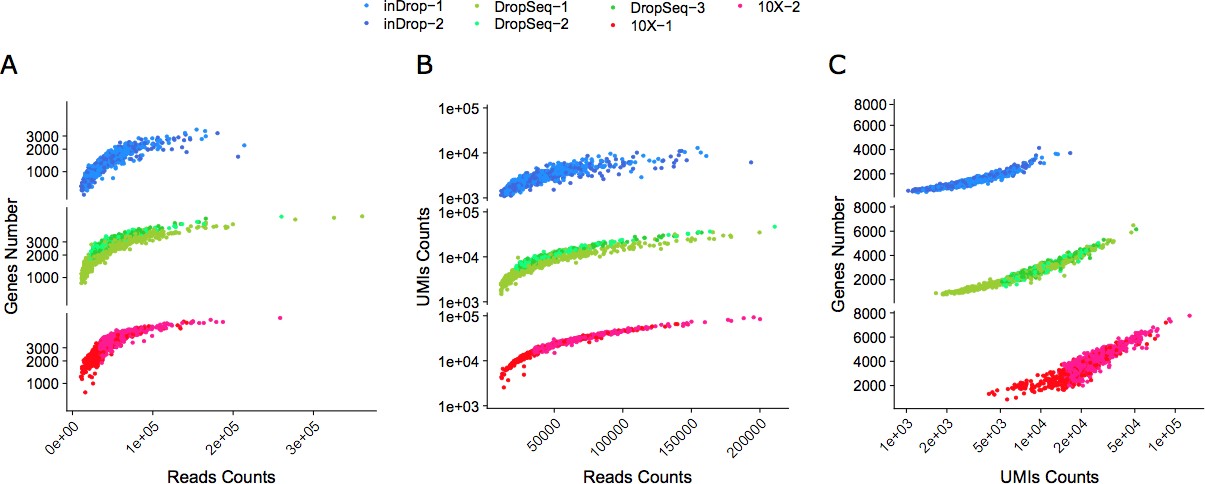


图S2(与图3相关)。数据处理管道的有效性。(A)我们比较了官方分析管道和我们的管道之间约100个细胞的聚集基因表达水平(log10转化)。(B)由官方管道和我们的管道量化的每个单元格中的umi(log 10-transformed)总数显示出高度一致性。(C)使用模拟数据比较umi的实际数量和我们的管道估计的数量。

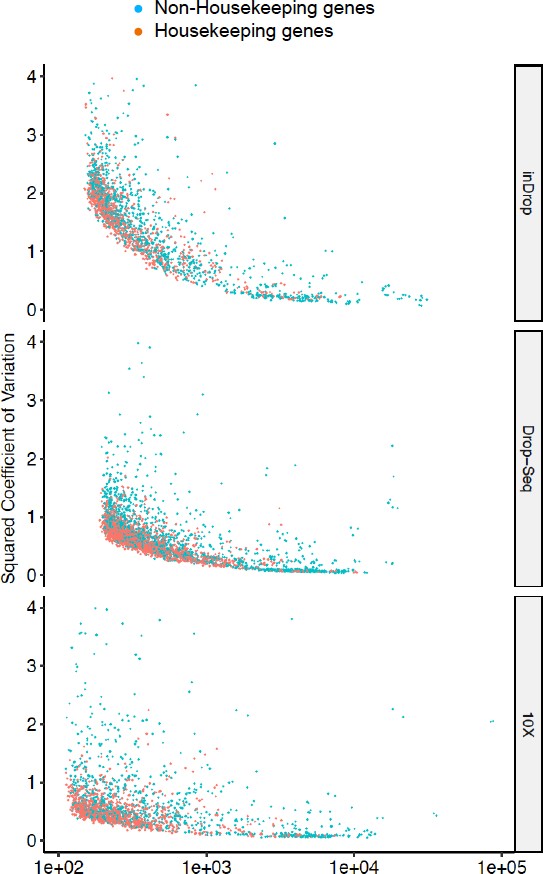


图S3(与图3相关)。确定有效的单元格编号。(A)条形码等级与读取计数图的推导；在估计的细胞数量周围可以看到悬崖。这对应于条形码等级与读取计数图中的拐点。用虚线(B)标记通过这种方法(称为“读取拐点法”)估计的细胞数。通过不同方法估计的细胞数(参见“方法”)。这里，术语“估计”表示通过考虑以下因素获得的数字

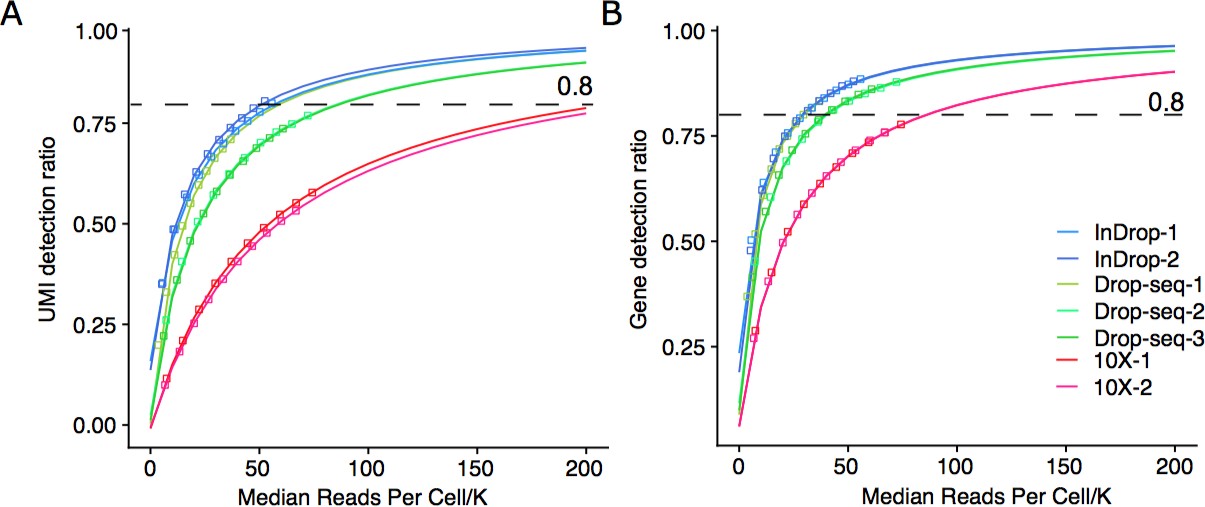
所有的实验证据。UMI (1/10)意味着UMI的较低级别被设置为条形码前1%级别的十分之一。读取≥20K或40K将需要相应的最小读取次数。



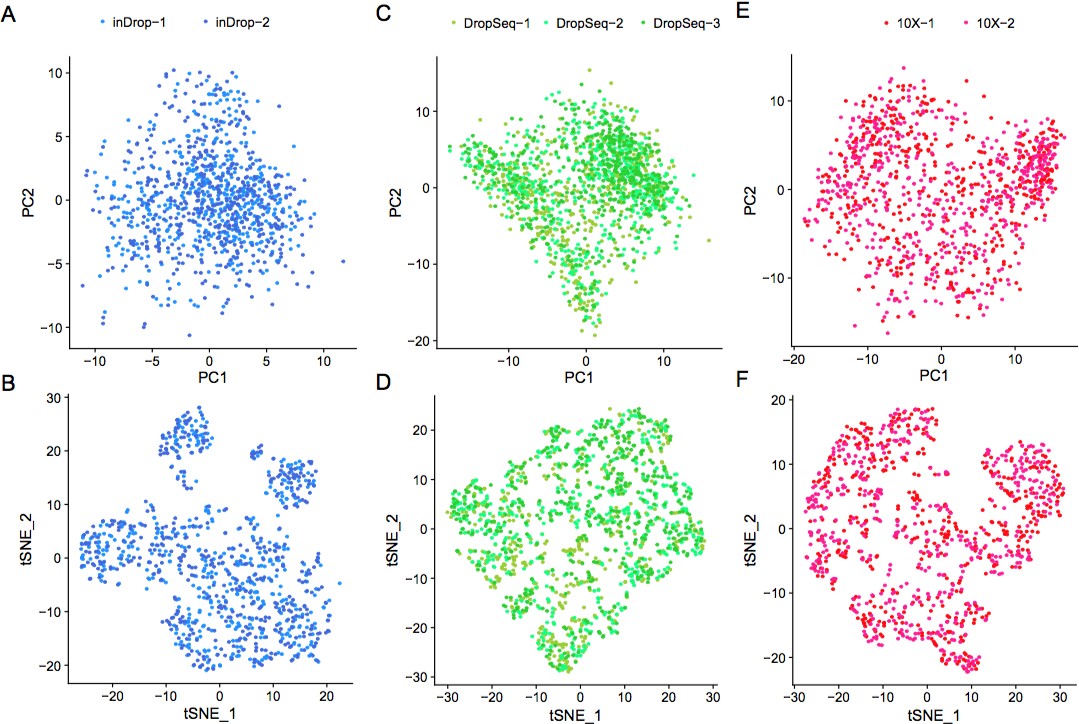
**图S4(与图4相关)。同一样本中所有细胞的基因表达灵敏度和精确度测量值之间的成对关系。(A)基因数量与阅读计数。(二)UMI计数与阅读计数。(三)基因数与UMI计数的关系。**



图S5(与图4相关)。每种方法的技术和生物噪声水平。管家和非管家基因用不同的颜色标记。它们分布的相似性表明技术噪声的水平超过了生物噪声。



图S6(与图5相关)。测序深度饱和度分析。绘制了每个二次取样测序深度时的UMIs)和基因(B)检测比。二次采样深度以样本中所有有效像元的中位读取数来测量。拟合曲线，预测饱和度。将检测比归一化为每个样本中UMI饱和水平或基因数(参见方法)。



图S7(与图6相关)。使用PCA和tSNE分析，使用相同系统对样本中的细胞进行聚类分析。系统为inDrop (A，B)、Drop-seq (C，D)和10X (E，F)。降维方法有PCA (A，C，E)和tSNE (B，D，F)。

表S1(与图2相关)。每个系统中检测到的前20个UMI序列。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| inDrop |  |  | 拖放序列 |  |  | 10X |  |
| UMI序列 | 深度 |  | UMI序列 | 深度 |  | UMI序列 | 深度 |
| CCCCCC | 20 |  | TTTTTTTT | 36.6 |  | tttttttttttt | 44.8 |
| ccccccct | 9.7 |  | GGGGGGGG | 17.9 |  | GTTTTTTTTT | 14.7 |
| ccccccca | 7.9 |  | CTTTTTTT | 13.1 |  | ATTTTTTTTT | 14.7 |
| TTTTTT | 7.1 |  | AGGGGGGG | 11.7 |  | CTTTTTTTTT | 12 |
| GCCCCC | 6.8 |  | GAGGGGGG | 11.3 |  | GGTTTTTTTT | 9.2 |
| ACCCCC | 6.8 |  | GGGGGGGT | 10.7 |  | TATTTTTTTT | 7.2 |
| CCCCCG | 6.5 |  | CGGGGGGG | 10.5 |  | AATTTTTTTT | 7.1 |
| CCCCTT | 6.1 |  | ATTTTTTT | 10.4 |  | GCTTTTTTTT | 七 |
| ATTTTT | 6.1 |  | GGAGGGGG | 10.2 |  | GATTTTTTTT | 6.9 |
| CCGCCC | 6 |  | TGGGGGGG | 10.1 |  | TGTTTTTTTT | 6.8 |
| AATTTT | 5.8 |  | GGGGGAGG | 10 |  | agtttttttttt | 6.6 |
| CCCGCC | 5.7 |  | GTTTTTTT | 10 |  | CATTTTTTTT | 6.4 |
| CCCACC | 5.6 |  | GGGGGGGA | 10 |  | TCTTTTTTTT | 6.2 |
| CGCCCC | 5.5 |  | GGGAGGGG | 9.9 |  | ACTTTTTTTT | 5.9 |
| CCACCC | 5.4 |  | GGGGAGGG | 9.3 |  | CGTTTTTTTT | 5.8 |
| AAATTT | 5.4 |  | GTGGGGGG | 8.7 |  | CCTTTTTTTT | 5.4 |
| CACCCC | 5.3 |  | GGGGGGAG | 8.7 |  | GGATTTTTTT | 5.2 |
| ccccccac | 5.3 |  | TCTTTTTT | 8.7 |  | GTATTTTTTT | 4.9 |
| CCCTCC | 5.2 |  | CCTTTTTT | 8.6 |  | agggttttttt | 4.7 |
| AAA级 | 5.1 |  | AAGGGGGG | 8.6 |  | GGCTTTTTTT | 4.7 |

归一化深度的计算方法如方法所示。

表S2(与STAR方法相关)。读-基因标记模拟中的误标记基因。

|  |  |
| --- | --- |
| 真实基因 | 错误标记为 |
| RPL41 | AC090498.1 |
| IGHA1 | IGHA2 |
| RPS15A | RP11-1035H13.3 |
| IGLC3 | IGLC2 |
| 气象情报(Meteorological Information) | AP 000 03 50.10 |
| RPL36A | RPL36A-HNRNPH2 |
| IFITM1 | IFITM2 |
| ARPC1B | ARPC1A |
| C20orf24 | TGIF2-C20orf24 |
| TUBA1B | TUBA1C |
| RPL36 | HSD11B1L |
| PSMA1 | 钠 |
| FKBP1A | FKBP1C |
| NDUFA11 | AC024592.12 |
| AC090498.1 | RPL41 |
| PA2G4 | RP11-603J24.9 |
| HNRNPA1 | HNRNPA1L2 |
| HNRNPA1L2 | HNRNPA1 |
| VAMP2 | RP11-599B13.6 |
| NDUFB8 | RP11-411B6.6 |
| FKBP11 | AC073610.5 |
| NDUFB8 | 钠 |
| TMSB4X | 钠 |
| TUBA1B | TUBA1A |

在读至基因标记模拟中标记错误的基因(参见方法)。这些基因大多在基因组中重叠或具有相似的序列。

表S3(与图4相关)。每个系统的高度可变基因。这些基因按方差系数(CV)排序。过滤标准如方法所示。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 10X |  |  |  | 拖放序列 | |  |  | inDrop |  |
| 基恩 | 履历 | Mean\_Norm\_UMI |  | 基恩 | 履历 | Mean\_Norm\_UMI |  | 基恩 | 履历 | Mean\_Norm\_UMI |
| IGLC3 | 2.551 | 2264.403 |  | HIST1H4C | 1.352 | 3083.957 |  | IGLC2 | 2.236 | 5811.328 |
| JCHAIN | 1.950 | 4792.352 |  | IGLC2 | 1.328 | 17186.632 |  | IGHM | 2.120 | 7665.682 |
| IGHA2 | 1.401 | 2186.472 |  | IGHM | 1.123 | 17755.828 |  | MALAT1 | 1.435 | 3064.135 |
| IGHA1 | 1.394 | 25045.873 |  | HSPA5 | 0.992 | 2918.335 |  | MTRNR2L12 | 1.082 | 6777.936 |
| IGLC2 | 1.364 | 105380.787 |  | MALAT1 | 0.963 | 10510.735 |  | MT-ATP8 | 0.917 | 2413.889 |
| FTH1 | 0.853 | 4607.734 |  | FTH1 | 0.891 | 3429.481 |  | FTH1 | 0.907 | 3365.871 |
| ISG15 | 0.852 | 2200.911 |  | 联邦长途电信实验室(Federal Telecommunications Laboratory) | 0.821 | 3181.098 |  | HLA-DRB1 | 0.892 | 2431.433 |
| ACTB | 0.850 | 7805.019 |  | HSP90B1 | 0.746 | 4117.068 |  | PTMA | 0.815 | 3285.668 |
| 联邦长途电信实验室(Federal Telecommunications Laboratory) | 0.811 | 6396.760 |  | MT-ATP8 | 0.713 | 2108.570 |  | SRGN | 0.798 | 6019.020 |
| ACTG1 | 0.765 | 2816.229 |  | ACTB | 0.692 | 4465.837 |  | HLA-DRA | 0.756 | 3967.253 |
| TMSB4X | 0.707 | 33627.828 |  | PDIA3 | 0.685 | 2405.884 |  | MT-ND5 | 0.756 | 5430.640 |
| PTMA | 0.705 | 4846.445 |  | MT-ND5 | 0.659 | 2831.306 |  | IFITM2 | 0.688 | 2359.603 |
| GAPDH | 0.696 | 5946.673 |  | HSP90AA1 | 0.563 | 2780.779 |  | PFN1 | 0.685 | 2810.210 |
| CD74 | 0.694 | 4239.958 |  | MT-ATP6 | 0.556 | 5223.957 |  | CD74 | 0.664 | 4899.420 |
| PFN1 | 0.653 | 2963.908 |  | HSPD1 | 0.545 | 2797.749 |  | ACTB | 0.659 | 16324.214 |
| MT-  ATP6 | 0.635 | 2122.471 |  | CYB山 | 0.521 | 2809.355 |  | HLA-C | 0.652 | 2820.471 |
| TMSB10 | 0.605 | 4284.120 |  | MT-ND2 | 0.516 | 4022.501 |  | RPS13 | 0.649 | 3211.120 |
| MALAT1 | 0.603 | 14898.154 |  | MT-ND3 | 0.514 | 2345.705 |  | HLA-B | 0.647 | 2862.303 |
| YBX1 | 0.588 | 2171.211 |  | HNRNPA2B1 | 0.510 | 2623.687 |  | NPM1 | 0.646 | 2572.229 |
| RPS24 | 0.566 | 2671.524 |  | MT-ND4 | 0.510 | 3831.651 |  | 人类白细胞抗原-A | 0.637 | 2451.213 |
| HLA-B | 0.553 | 2257.010 |  | MT-CO2 | 0.506 | 5319.697 |  | GAPDH | 0.631 | 2739.710 |
| MT-CO3 | 0.544 | 7057.905 |  | MT-ND1 | 0.504 | 4194.428 |  | CYB山 | 0.609 | 10615.063 |
| MT-ND2 | 0.542 | 3116.763 |  | NCL | 0.503 | 3348.467 |  | PPIA | 0.602 | 2400.896 |
| CYB山 | 0.542 | 3246.955 |  |  |  |  |  | TMSB4X | 0.591 | 15151.985 |
| SUB1 | 0.530 | 2086.202 |  |  |  |  |  | RPL39 | 0.573 | 2512.036 |
|  |  |  |  |  |  |  |  | RPS20 | 0.562 | 2539.395 |
|  |  |  |  |  |  |  |  | RPS10 | 0.560 | 2131.890 |
|  |  |  |  |  |  |  |  | UBA52 | 0.560 | 2410.621 |
|  |  |  |  |  |  |  |  | EIF1 | 0.557 | 4003.951 |
|  |  |  |  |  |  |  |  | MT-ATP6 | 0.557 | 15550.839 |
|  |  |  |  |  |  |  |  | ACTG1 | 0.556 | 5873.674 |
|  |  |  |  |  |  |  |  | MT-ND3 | 0.554 | 17580.354 |
|  |  |  |  |  |  |  |  | RPL19 | 0.522 | 2308.035 |
|  |  |  |  |  |  |  |  | RPLP1 | 0.520 | 4997.513 |
|  |  |  |  |  |  |  |  | RPL14 | 0.512 | 2590.718 |
|  |  |  |  |  |  |  |  | MT-ND2 | 0.512 | 17652.350 |

表S4(与图7相关)。各种条件下inDrop系统中Smart-seq2四次实验的基因和UMI检测性能统计。

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 原始读取 | 有效读取次数(白名单中有条形码) | 索赔的单元格 | 已声明的单元格读取 | 声称拥有细胞UMI读取 | 有效CB  计数 | 中位读数 | 中位UMI-  有效读取 | 基因计数中位数 | 中位UMI  计数 |
| 腰神经2 | 203213268 | 123,672,238 | 1300 | 100,595,865 | 61,688,856 | 366 | 101543 | 62509 | 1075 | 2939 |
| L2 | 257489533 | 165,149,928 | 900 | 124,648,424 | 80,839,124 | 492 | 108251 | 70966 | 822 | 1949 |
| S1 | 293336617 | 182,604,144 | 1,300 | 152,093,346 | 100,414,242 | 440 | 113533 | 76759 | 925 | 2560 |
| S2 | 221840332 | 144,839,268 | 650 | 123,550,924 | 80,182,371 | 176 | 101817 | 66333 | 996 | 2841 |

L1是逆转录+池聚合酶链反应。L2是42/50℃RT+裂解PCR。S1为42 ℃，RT+Pool PCR。S2是42 ℃逆转录+裂解聚合酶链反应。L1和L2的逆转录反应温度为42/50 ℃，S1和S2为42 ℃。在S1和L1，cDNA被汇集在一起进行扩增。在S2和L2，cDNA被裂解用于PCR。

表S5。(与图4相关)三个系统中一些已发布数据集的汇总。证明了组织或细胞系、细胞数量、测序深度和检测到的umi。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 草案 | 单元格类型 | 来源 | UMI >= 1000哥伦比亚广播公司 | 中位umi | 每次读取总次数  细胞 | 每个的映射读取  细胞 |
| inDrop | HEK293 | 我们的\* | 2,500 | 6,373 | 39,189 | 35,034 |
| inDrop | 斑马鱼14hpf | 3 | 4,001 | 7,259 | 53,379 | 37,831 |
| inDrop | 鼠标ESC | 一 | 703 | 5,667 | 40,088 | 34,119 |
| inDrop | 人类Treg | 2 | 1,726 | 1,647 | 32,774 | 20,217 |
| 拖放序列 | HEK293 | 四 | 587 | 11,472 | 32,243 | 28,771 |
| 拖放序列 | 斑马鱼12hpf | 5 | 4,413 | 2,497 | 钠 | 11,945 |
| 拖放序列 | 小鼠视网膜 | 四 | 3,207 | 4,398 | 钠 | 钠 |
| 10X | HEK293 | 8 | 504 | 21,400 | 32,361 | 31,260 |
| 10X | HEK293\* | 我们的\*\* | 2,288 | 9,015\* | 16,321\* | 15,522 |
| 10X | 斑马鱼12hpf | 5 | 3,000 | 10,577 | 钠 | 钠 |
| 10X | 小鼠脑 | 6 | 931 | 8,545 | 56,718 | 53,372 |
| 10X | 人T细胞 | 七 | 3,465 | 3,731 | 31,788 | 29,709 |

参考:

1. [1]克莱因，马祖蒂斯，阿卡尔图纳，等.应用于胚胎干细胞的单细胞转录组学液滴条形码[J].中国农业大学学报，2003，22(3).Cell，2015，161(5): 1187-1201。
2. Zemmour D，Zilionis R，Kiner E，等.单细胞基因表达揭示了一个由TCR决定的调节性T细胞表型景观[J]。《自然免疫学》，2018，19(3): 291。
3. [1]Wagner D E，Weinreb C，Collins Z M，等.斑马鱼胚胎基因表达谱和谱系的单细胞定位[J]。《科学》，2018，360(6392): 981-987。
4. Macosko E Z，Basu A，Satija R，等.使用纳升液滴对单个细胞进行高度平行的全基因组表达谱分析[J]。Cell，2015，161(5): 1202-1214。
5. [1]法雷尔，王勇，李森菲尔德，等.斑马鱼胚胎发生过程中发育轨迹的单细胞重建[J].中国水产大学学报，2003，25(2).《科学》，2018，360(6392): eaar3131。
6. <https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/neurons_900>
7. <https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/t_3k>
8. <https://support.10xgenomics.com/single-cell-gene-expression/datasets/2.1.0/hgmm_1k>

符号:

\*不同批次的珠子。

\* \*我们使用了一半的cDNA进行下游文库制备和测序(参见方法)。

表S6。(与STAR方法相关)所有三种平台每细胞RNA-seq文库制备成本的估算。

|  |  |  |  |
| --- | --- | --- | --- |
|  | inDrop | 拖放序列 | 10X |
| 仪器价格 | 50,000美元 | 30,000美元 | 50,000美元 |
| 每个单元的仪器成本1 | $0.07 | $0.04 | $0.07 |
| 每个单元的运行成本 | $0.10 | $0.10 | $0.50 |
| 每个细胞的测序成本2 | $0.30 | $0.30 | $0.30 |
| 每个单元的总成本 | $0.47 | $0.44 | $0.87 |

符号:

1:我们假设每个平台每天制备1000个细胞，每个仪器可以提供2年的服务。

2:假设每个细胞平均读取50000次(150bp对端)。