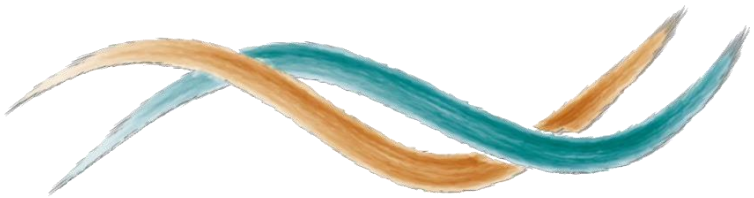
10x 单细胞转录组常见问题



安诺优达基因科技（北京）有限公司

单细胞业务线

2022/3/11

目录

一、建库测序环节

1.细胞活性检测、荧光染料

AO/PI 双染细胞凋亡检测试剂盒。
吖啶橙(Acridine Orange，AO)可透过正常细胞膜，使细胞核呈绿色或黄绿色均匀荧光；而在凋亡细胞中，因染色质固缩或断裂为大小不等的片断，形成凋亡小体。吖啶橙使其染上致密浓染的黄绿色荧光，或黄绿色碎片颗粒；而坏死细胞黄荧光减弱甚至消失。
碘化丙啶（Propidium Iodide，PI），无膜通透性，不能透过活细胞膜，只能染死细胞。因此，在荧光显微镜下观察，正常细胞不能着色，早期凋亡细胞呈微弱红光，晚期凋亡细胞红光加强，坏死细胞呈强红色荧光。
双染料进行染色的情况下，活细胞呈现绿色或黄绿色均匀荧光；死细胞呈现明显的红色；而活性较差、开始凋亡的细胞颜色在两者之间，通常是黄色、橙色等。

2.样本准备时，如何降低背景 RNA?

样本制备过程，细胞破碎会导致细胞悬浮液中游离 mRNA 存在，同样可能会被 10x 分选捕获，导致低质量异常数据，因此需要降低背景 RNA。
(1) 上样前，用 buffer 清洗 2 遍单细胞悬浮液
(2) 尽量缩短上样时间，以防止细胞裂解。
(3) 理想情况下，细胞存活率应为 90%，以减少来自裂解细胞的总 RNA 量。

3.样本上机前是否应该去除红细胞？

不是必须的，但如果实验目的不关注红细胞，则尽量去除红细胞，避免不必要的数据量浪费，否则会因红细胞的存在导致测序数据量的增加。红细胞去除方法有很多，例如官方肿瘤解离 protocol 中有提到红细胞裂解,同样也有老师用 Ficoll 去除 PBMC 中的红细胞。
注：红细胞特异性转录本在 cDNA 质检中可能出现约 700 bp 的尖峰。这是正常的，不会影响下游文库的质量。

4.影响捕获细胞数目的因素有哪些？

主要影响细胞捕获数目多少的是样本处理成单细胞悬液的过程——细胞浓度与活性。样本物种类型及组织特点多样，如骨组织，灌洗液，成熟心肌组织或植物原生质体等，需要较为特殊的处理方式，以保障细胞活性；外周血单核细胞分离时需要注意避免其他细胞的污染等。单细胞研究板块致力于样本制备，目前已成功测试过脾脏、肾脏、乳腺癌原位灶及转移灶、人-小鼠移植瘤、小鼠结肠癌原位灶与结直肠等。

5.样本中的 RNA 含量低的原因大概有哪些？

原因如下：
1）依据细胞类型考虑，与大型细胞比，小型细胞趋向于 RNA 含量低。当然，大型细胞或细胞核样本也会存在 RNA 含量低。通常原代细胞和细胞直径小的细胞（如 PBMC）的 RNA 含量低于细胞系或癌细胞。
2）单细胞悬浮液制备：

a 样本准备过程：由于解离和 FACS 等导致细胞裂解；
b 细胞活性：当细胞进入凋亡状态时，较低的细胞存活率会降低转录输出；
c PCR 循环数：如果所选细胞为小型细胞或细胞核等 RNA 含量低的细胞，则在 cDNA 扩增中和/或样本 index PCR 步骤中增加额外的 1-2 个 PCR 循环。此时需注意， 额外的 PCR 循环可能增加 PCR 产品，但可能不增加最终文库复杂度。

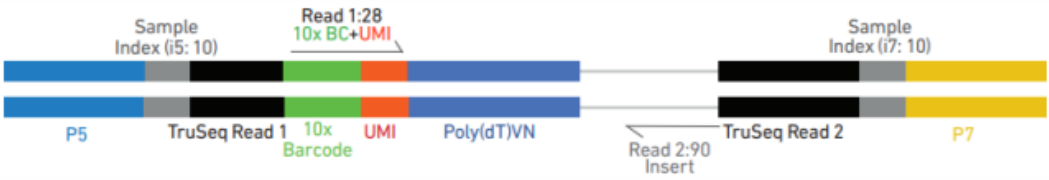
6.影响 cDNA 浓度的主要因素有哪些？

（1）样本的类型会影响扩增后的 cDNA 浓度，不同细胞的基因表达水平不一样，另外有些细胞容易粘黏或贴管壁。
（2）细胞捕获效率，直接影响细胞捕获效率。比如：细胞活性与浓度、细胞状态、组织解离、FACS 等。样本上芯片前的存放时间，实验操作是否严谨，活性浓度计数是否准确，及 PCR 循环数（增加 cDNA 的 PCR 扩增循环或样本 index 的循环数可能会导致文库 duplicates 比例升高）等。

二、数据分析环节

1.Clean data 和 Raw data 关系

一般 clean data 是 raw data 的 40%内。和普通转录组不同，目前 10x 单细胞转录组使用的试剂版本：Reagent V3.1
R1: 16bp Barcodes + 12bp UMI
R2: 90bp insert



2. 什么是测序饱和度（sequencing saturation）？

测序饱和度用于衡量本次实验下被测到的文库复杂度分数。测序饱和度这个概念可以侧面说明想要检测到一个新转录本所需额外 reads 数目。
影响测序饱和度的因素有文库复杂度和测序深度。不同细胞类型的 RNA 含量不同，所以最终文库中不同的转录本总数有差异，即文库复杂度不同。
如图所示，细胞类型不同，对应的基因检出中位数也不同，但基因检测中位数都随着测序深度的增加而增加，且不同细胞类型达到饱和时的测序深度均不相同。
因此，测序深度越高，检测到额外的新转录本越多，然而，这又限制于文库复杂度。同等测序数据量下，复杂度低的文库一般会具有更高的饱和度。
网页版介绍：https://www.sohu.com/a/415091996_120084361

3. 测序饱和度应该达到多少呢？

测序饱和度是测定测序中捕获到的文库复杂度的分数。依据实验目的的不同，从而判断是否需要高的测序饱和度。

如果您只是关注细胞分群及其下游分析，不需要检测到每个细胞中的每个唯一转录本，则较低的饱和度就足够；如果您试图捕获低表达的转录本，则需要较高的饱和度。原代细胞（如 PBMCs）的 RNA 含量比较低，可能较低的测序深度即可获得高的测序饱和度（90%以上）。

测序饱和度侧面可以说明想要发现一个新的 read 能检测到新转录本的概率。如果测序饱和度为 50%，这意味着每 2 个新 read 中将检测到一个新 UMI count（唯一的转录本）。相反，测序饱和度为 90%，则意味着 10 个新 read 中只能检测到 1 个新 UMI count。如果测序饱和度很高，额外的测序（加测）将很难发现文库中更多的信息。

4. 为什么样本中有线粒体基因出现高表达的情况?

A、样本质量差，从而导致大量细胞凋亡或降解。

B、特定样本的生物学特性，例如肿瘤活检(穿刺样本)，由于代谢活动或细胞坏死，可能增加了线粒体基因表达。

(1)凋亡细胞表达线粒体基因，并将这些转录本输出到哺乳动物细胞的细胞质中。当凋亡细胞被加入到其他正常细胞悬浮液中时，导致后期检测到的线粒体基因数会增加。

(2)裂解细胞或细胞膜被破坏的细胞释放其胞质转录本，而线粒体转录本可能仍保留在双膜结合的线粒体内。线粒体完整的裂解细胞可能被包进 GEMs 中，从而导致检测到的线粒体 转录本比例增加。

(3)如果仅 1 个或几个 cluster 有线粒体转录本差异上调和较低的总 UMI count，那么这个 或几个 cluster 最有可能代表死亡或垂死细胞群。

5.对重复样本进行合并分析的时候，使用的是什么标准化的方法?

答：我们采用 cellranger aggr 模块进行多个样本的合并；不同样本之间的数据标准化的方法是采用 cellranger aggr 的参数 mapped。

mapped (default): For each library type, subsample reads from higher-depth GEM wells until they all have an equal number of reads that are confidently mapped to the transcriptome or assigned to the feature IDs per cell.

网页版的介绍： https://kb.10xGenomics.com/hc/en-us/articles/115004217543-How does-cellranger-aggr-normalize-for-sequencing-depth-among-multiple-libraries-

6.分析中具体的均一化方法是什么？？

答：seurat 分析中，默认采用 LogNormalize 归一化算法。该归一化对低表达的基因没有影响

7.差异基因

1）差异基因判断标准是什么

答：差异基因表达的判断标准是 logFoldChange，非 P 值， K-mear 值，seurat 软件结果只对 avg_logFC 有个阈值控制（seurat 软件默认），一般为 0.25，对其他值比如 P_val 和 p_val_adj 都没有设定阈值，所以会出现有 p_val_adj 值为 1 的结果。主要原因

是由于单细胞数据表达量数据较低（与 bulk RNA 相比较），设定太严格的阈值可能会造成有些有意义的数据被过滤掉了

2） 差异 marker 比较组是怎么设定的，如何判断是上调还是下调？

答： Seurat analysis 里单个 cluster 中的 marker 基因的上下调显著结果是与其他 cluster 中该基因的平均表达量相比较得到的结果。

3） 结果文件夹中个别项目所给的 top 1 或者 top10 基因，与 csv 电子表格中所呈现的基因并不是对应，是根据什么选择的 top1 或者 top10 基因呢？

答：csv 电子表格所排序是按照 p_val_adj 进行排序的，我们作图的时候是按 avg_logFC 进行排序。一般认为 avg_logFC 值越大，其差异也越大。

4）avg_logFC 数值有有正有负，依据什么原则进行排序选择呢？

答：avg_logFC 值的正负可以认为看做是上下调关系，一般值为正，其为上调，其值为负为下调，一般看结果的时候，我们优先看正值，即上调基因。

三、售后处理环节

1.细胞注释相关问题

1）如何寻找细胞 marker 基因，用于细胞注释

要想对细胞进行定义，我们就需要找到可以定义这些细胞的已知的 marker genes，所以 marker genes 的寻找也就成为定义细胞群的关键。

寻找 marker genes 的方法推荐：

a 通过阅读文献来寻找，这个一般也是比较可靠的；

b 通过 cell marker 的一些网站进行寻找，下表网站供参考；

数据库	物种	网址
MCA	小鼠	http://bis.zju.edu.cn/MCA/index.html
cellmarker	人，小鼠	http://biocc.hrbmu.edu.cn/CellMarker/index.jsp
panglao	人，小鼠	https://panglaodb.se/index.html
CancerSEA	肿瘤	http://biocc.hrbmu.edu.cn/CancerSEA/goDownload

c 可以通过分析结果中的特异 marker genes 来判断，这个需要经验丰富，或者查阅资料，确定这些 marker genes 的功能以及主要存在什么细胞类型里面；

d 基于已公布的转录组数据进行相关性分析以确定细胞类型。

对于已知 marker genes 定义不了的群体如何注释：

a 将自己单细胞数据的 clusters 中的基因表达与先前发表的单细胞转录组进行了比较并建立回归模型，对细胞群进行了定义，这种方法可以识别独特的单细胞簇；

b 通过与现有的 bulk 转录组数据比较，进行相关性统计，通过相关性大小来区分未知群与已知群的相关性以达到定义细胞的目的。

参考文献：Kristofer D，Jasper J，Duygu K，et al. A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain[J]. Cell, 2018:S0092867418307207.

2）是否可以根据 LogFoldChange 降序选择某 cluster 前五个 marker 基因来确定细胞类型？

答： 对分群得到的不同细胞分类进行细胞类型确定，可以根据上述结果中不同 cluster 中的 marker 基因来确定

2.PCA 主分析的数量为什么与细胞簇不一致

答：PCA 分析是一种数据降维手段，即将 n 维特征映射到 k 维上，这 k 维是全新的正交特征也被称为主成分，PCA 分析得到的 12 个 PCA 是可以解释这些变量的主成分，不是细胞的聚类，不对应细胞。

3.同一个样本的三个基本数据 D1.barcode.tsv D1.genes.tsv D1.matrix.mtx ，使用 read10x 指令时显示 barcode file missing，使用 read table 指令是也无法读出，请 问是文件有问题吗？

答：Read10x 载入数据时，路径下的名字需要为这三个文件（"barcodes.tsv" "genes.tsv" "matrix.mtx"），不能带有样本名称 ，修改名称去掉样本名称前缀，重新尝试。

4.在数据分析过程中,利用 SeuratR 包分析的中间数据(.rds 格式)，是否有保 存，能否提供给客户？能否利用 SingleR 包，进行全自动注释？

答： rds 格式的文件可提供给客户，可向项目管理负责人要；SingleR 包没有在我们的 10x Genomics 分析流程中， 用户可输入 rds 文件和对应的结题报告中的 upload/1_CellRanger_analysis/2_expression/*all_UMI.csv 文件，即可运行

5.Reads Mapped Confidently to Transcriptome < 60%怎么办？

该指标一般适用于人或小鼠等模式生物，其他物种该指标不一定适用。
对于一些特殊组织，也不一定适用，如，一般造血干细胞、细胞核与胚胎组织等，未剪切的 RNA 含量较高，因此内含子比对会偏高，而外显子比对偏低。建议用前体 RNA 做参考基因组，再进行比对，可能会提升基因数。
前体 RNA 参考基因组为将 GTF 文件中第 3 列替换成外显子后再次生成的参考基因组。
相关链接：
https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/advanced/references - header

6.为什么相同的数据 Cell Ranger 分析后，得到的 t-SNE 图会不一样？

t-SNE 图是展示对细胞聚类一种展示形式,与 PCA 不同,t-SNE 的非线形降维是不确定的，但这不影响数据分析结果。10x 官方回复：“In contrast to, e.g., PCA, t-SNE has a non-convex objective function. The objective function is minimized using a gradient descent optimization that is initiated randomly. As a result, it is possible that different runs give you different solutions.”
Thus subtle differences like you observed between runs are not a concern. For more information please see: https://lvdmaaten.github.io/tsne/

7.Reads Mapped to Genome 明显低于常规水平？

除样本中有不同物种细胞混合外，一般原因为物种与参考基因组不匹配。

8.影响基因检出的因素有哪些？

（1)主要与样本类型及状态有关,一般外周血多为 1000~3000,干细胞和大脑可多达 5000+，不同的癌种，基因检出数也可存在数倍的差异；

（2）随着测序饱和度的增加，在一定程度上会增加基因检出数，一般测序饱和度达到 80%即可，但并不是硬性指标；

（3）参考基因组的注释是否完善会影响基因检出，若有特殊物种，则需要提供较为完善的参考基因组。

（4）对于一些特殊组织，如，一般造血干细胞、细胞核与胚胎组织等，未剪切的 RNA 含量较高，因此内含子比对会偏高，而外显子比对偏低。建议用前体 RNA 做参考基因组，再进行比对，可能会提升基因数。