**单细胞转录组分析与质控基本要点**

张霈 2022/9/23

1. **为什么要做单细胞转录组？**

有机体的各种器官或组织，有不同类型的细胞组成，它们的协作分工是器官发挥功能的关键。因此，获得各种细胞类型的转录特征，对我们研究各种组织器官的运作至关重要。普遍意义上的“单细胞转录组”可以分为捕获细胞质成熟mRNA的真正单细胞转录组，和捕获细胞核中未成熟mRNA的单细胞核转录组。一般认为，单细胞转录组测序质量更高；但由于细胞形态的多样化，某些时候，单核转录组可能是更好的选择。**整体上，基于表达量分析的单细胞转录组和单核转录组的分析思路不会有太大差异，但部分质控标准需要改变**。

对于发育和癌症等涉及细胞状态改变的研究来说，RNA velocity 同样是非常重要的研究对象。RNA velocity，指单个基因的nascent mRNA和 mature mRNA的相对比例。我们可以假设，在一个稳定的细胞中，关键基因的mRNA转录、剪切、降解的速度基本恒定，因此RNA velocity也是恒定。而如果细胞状态发生改变，那么自然关键基因的RNA velocity也会发生改变。因此，RNA velocity同样也可以用于细胞识别和分群。目前很多方法都用到了RNA velocity信息，如2022年2月发表在《Nature methods》上的Cellrank方法等。若要使用RNA velocity，原则上**单细胞和单核测序**均可以，但需要预先评估un-spliced ratio。目前主要方法和模型还是基于单细胞测序开发的（虽然原则上单核测序有更高的un-spliced ratio）, 所以原则上应该使用**单细胞测序**。

1. **华大单细胞测序技术简述。**

类似Drop-seq（Figure. 1）, 利用微流控技术和barcode技术捕获单细胞。然后对捕获的mRNA ploy-A富集后3’端进行测序。因此，该技术并非全长转录组测序，只能进行reference guided的基于表达量相关的分析。更复杂的分析如单细胞可变剪切等均无法进行。

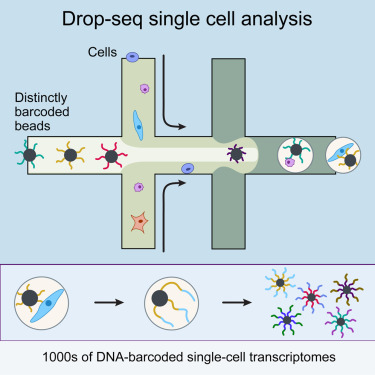


Figure. 1: Drop-seq示意图。

实验发现，普通Drop-seq中，很多液滴有细胞但无beads。 华大DNBelab C4 scRNA-seq v2平台为了更高效率的捕获细胞，会有意增加beads数目，可能导致一个液滴中有多个beads（多个beads捕获一个细胞）。为了在后续分析中合并属于一个液滴/细胞的beads, 华大平台spike-in M280 人工序列。

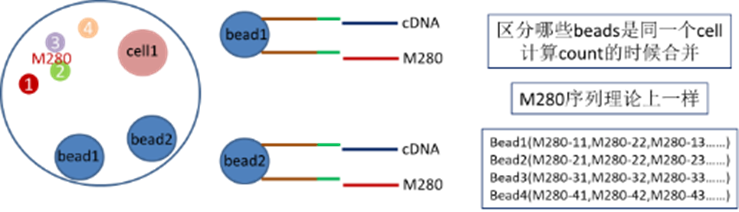


Figure.2: M280的spike-in与barcode.

1. **华大数据下机格式**

华大单细胞转录组下机数据包含两组fq文件：一组是pair-end cDNA测序结果，可能有一或多对fq.gz；另一组是pair-end M280测序结果。在cDNA测序结果中，fq1只包含barcode信息，包括20bp cell (bead) barcode信息+10bp UMI barcode信息；fq2则是和fq1对应的链特异性cDNA序列。所以，尽管cDNA测序结果包括一对fq文件，但严格上说是链特异性的单端cDNA测序。而在M280测序结果中，fq1是cell (bead) barcode信息, fq2是M280序列（可以类似为M280的UMI barcode）。

1. **单细胞转录组分析的前期准备**

如上述所说，该分析是reference guided, 所以项目开始前，必须保证研究对象有高质量的基因组组装和基因集的注释。考虑到Mitochondria的表达情况是评估细胞质量的重要参数，因此，reference 基因组和基因集中必须有确定无疑的线粒体DNA组装与注释结果，而且需要在ID或Name的开头加上可识别的字符串，以供后续分析识别。而基因注释集中，应该只包含有生物学意义的基因，主要的蛋白编码基因与lncRNA基因，以及用来评估细胞质量的线粒体基因，格式可以使用ensembl 标准gtf格式。在进行分析前，需要检测基因ID以及基因Name, 确保不会出现多个基因共享一个ID或Name的情况。我们使用Star基因cDNA比对，因此请用Star对基因组建立Index。

1. **华大基础分析流程**

目前华大基础分析流程由细胞所开发，在我们小组，由**朱元镇**（[zhuyuanzhen@genomics.cn](mailto:zhuyuanzhen@genomics.cn)） 整理，具体路径在*/hwfssz1/ST\_EARTH/P18Z10200N0107/P17Z10200N0101\_Metazoa\_RNA\_Editing/zhuyuanzhen/00bin/scRNA-seq-V2/scRNA-seq-v2.py* 如果由问题的话，可以咨询zhangpei@genomics.cn或朱元镇。

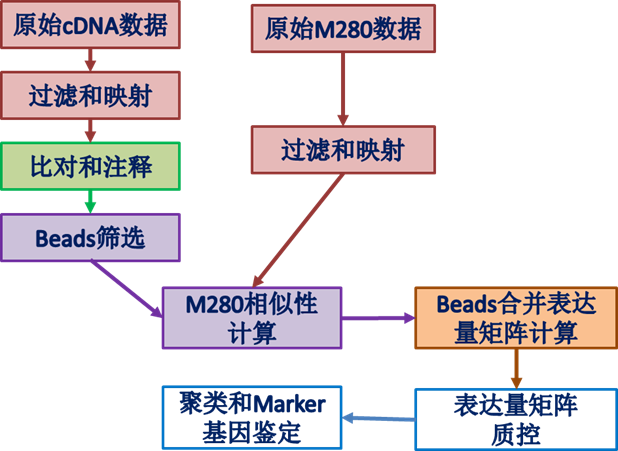


Figure 3. 华大单细胞转录组测序结果基本分析流程。

具体分析与质控：

**下机数据**，首先估算下机数据大小（测到的reads数）是否符合标准，然后进行基础质控，包括过滤掉低质量测序结果，过滤掉非正常的barcodes等。统计可用数据比例。目前过滤后，可用数据量大概占所有测序数据量75%以上。如果过低，该文库并不是完全不能用，但后续分析需要谨慎。

**cDNA比对。**华大流程使用Star进行比对，然后使用细胞所开发的PISA进行注释和UMI counts。如无意外，Star比对用默认参数即可，注意评估比对结果，包括比回率，unique比回率，junction reads比例，junction类型比例等。以上信息均在“**Log.final.out**”中可见。如果以上统计结果异常，则说明样品质量过差，必要时终止分析。

**比对结果映射。**华大使用自己开发的PISA将比对结果映射到基因注释结果中，根据测序类型，在PISA anno 这一步中选择不同参数，如果时单细胞测序，建议选择-splice-consider或-intron (测试发现基本不会影响结果)； 如果是单核测序，一定选择-intron.

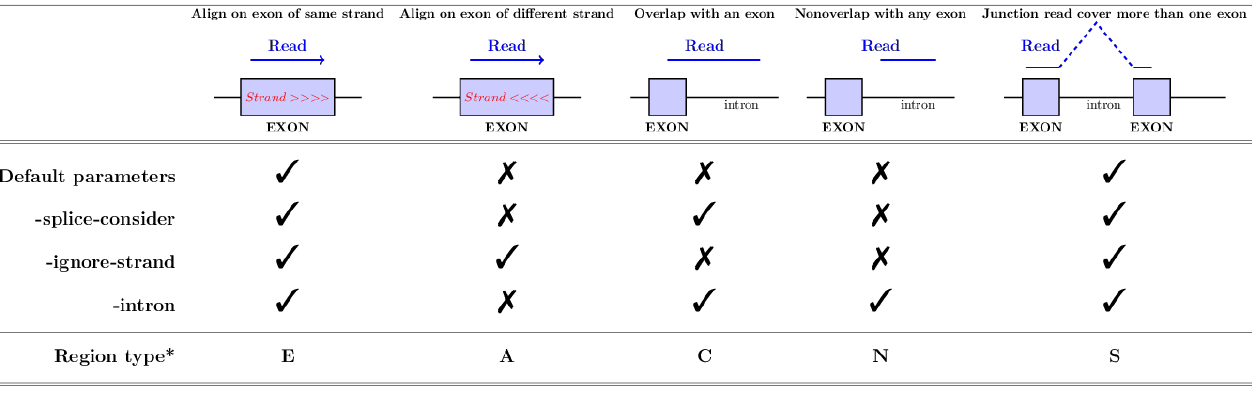


Figure 4. 需要根据测序类型手动设置的参数。

**PISA corr**是对UMI barcode进行校正，如果来自同一细胞同一基因的两个UMI barcode仅仅有1nt差异，则过滤掉测序次数低的那个。默认设置。之后是**PISA attrcnt**, 简单的数出每个beards测到的reads数，转录本数，基因数，为后续的beads过滤和比对结果过滤做准备。

**空液滴过滤**。华大流程使用空液滴过滤曲线进行beads的空液滴过滤。即将beads根据测到的转录本数据排序，然后画出beads rank – UMI数曲线，陡降的位置是阈值。目前该方法非常不灵活，需要根据结果手动调整阈值期望-e (期望细胞数，默认为7000)，目前华大技术指标为5000~10000. 如果有必要的话，对于单核测序，也可以使用Cellbender (<https://cellbender.readthedocs.io/en/latest/>) 选择阈值，程序具体路径为/ldfssz1/ST\_DIVERSITY/PUB/USER/zhangpei/bin/python3.7.13/bin/cellbender。

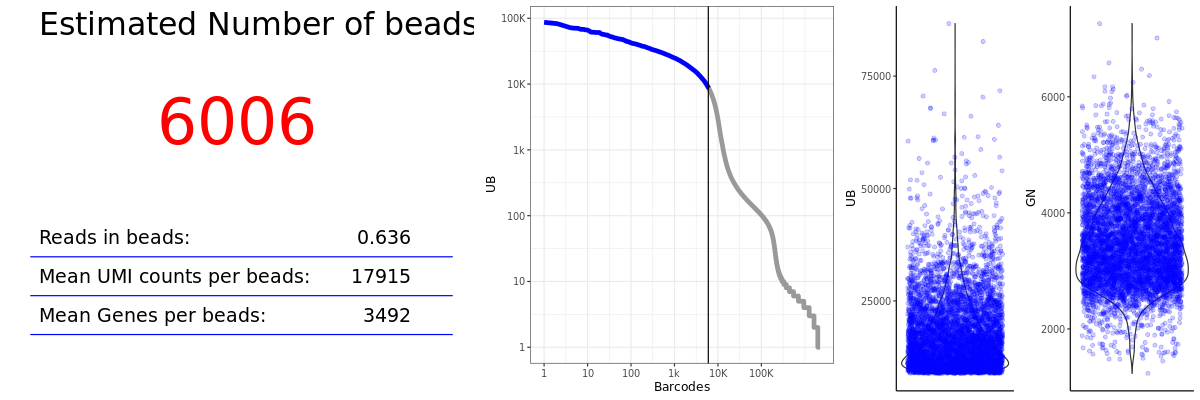


Figure 5. 空液滴过滤曲线结果示意。

**合并beads。**过滤掉空液滴的beads后，由于华大单细胞转录组策略，存在一个细胞被多个beads捕获的情况。因此，需要根据M280测序结果来进行合并。Beads在捕获细胞后，在液滴内，磁珠会释放M280序列，然后同样被beads捕获到。每一种M280序列可以视为一个基因，来自相同液滴的不同beads就会有相似M280 表达模式。根据M280测序结果得到每个M280序列在每个beads中的表达矩阵，然后计算beads之间的余弦相似性，然后余弦相似性大于0.5的beads合并。根据合并的beads，更新cDNA比对结果，然后使用PISA进行基因表达计数。

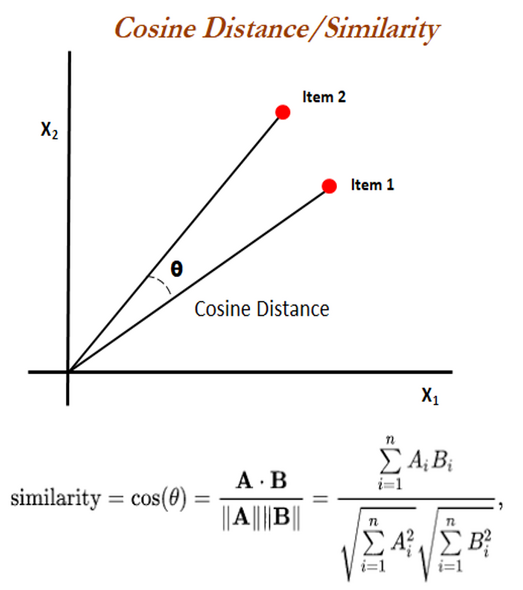


Figure 6. 余弦相似性，可以推广到高维空间。相对于传统欧几里得距离，更适合高维空间距离计算。

**最终结果**。生成10x格式的基因在每个细胞中的转录本表达数矩阵。同时给出一些基本质控信息，包括细胞数、每个细胞平均基因表达数、平均转录本表达数，以及线粒体表达比例等。如果细胞数、每个细胞平均基因表达数、平均转录本表达数过低的话，说明该文库质量较差。此外，每个细胞的线粒体表达比例是评估细胞质量的重要指标。通常来说，线粒体基因高表达的细胞一般是损伤或凋亡的细胞。对于单细胞测序，根据需要可以要求线粒体表达比例低于5%或10%；对于单核测序，一般要求线粒体表达比例低于1%。

1. **初步分析和质控。**

目前我们组主要通过Seurat V3分析平台进行单细胞分析，首先是对每个文库进行质控和前期处理。具体流程为：/ldfssz1/ST\_DIVERSITY/P18Z10200N0107/zhangpei/0.bin/1.scRNA\_seurat\_pipeline/1.seurat.QC.R，需要使用/hwfssz1/ST\_EARTH/Reference/ST\_DIVERSITY/PUB/USER/zhangpei/bin/R-4.1/bin/Rscript运行。如何有问题的话，[请联系zhangpei@genomics.cn](mailto:请联系zhangpei@genomics.cn)。

重要参数包括

--input：10x 格式转录本表达结果。

--mregex：线粒体基因模式匹配。

--mini：每个细胞表达基因数最小过滤阈值。

--maxi：每个细胞表达基因数最大过滤阈值，超过这个数字则被认为是双胞。

--mt.percent：线粒体表达比例，高于该值的细胞会被过滤掉，单细胞测序可以选择5或10，单核测序选择1。

--transform：表达量计算方式，建议选择sct。

--nf：高变基因数，根据经验和样品具体情况选择，一般是2000~5000。

--dim：PCA维度数，默认为50。

--dusage: 认为有生物学意义的维度数，先跑一遍，然后根据具体情况选择，对于初期质控来说，20就可以了。

--percentage：使用DoubletFinder评估合胞情况，然后过滤掉最有可能合胞的细胞，该参数绝大有多大比例细胞会被过滤掉。默认5%。

--res: 分群精度，初期质控默认就可以了。

--out：输出结果目录。

--marker：如果对分群结果有预期，给出已经知道的marker gene list, 有助于更好的评估分群质量，并非必要。

--cc\_gene：如果是胚胎测序，需要对细胞周期基因进行校正，如果设1的话，使用Seurat自带的cc genes。 也可以提供一个自己生成的cc genes list, 格式包括两列，基因name 和type (s.genes和g2m.genes)。非细胞分裂活跃样品，不要设置该参数。

--sample：样品、文库名，整个project中，每个文库的sample应该都是唯一。

具体流程是根据读取*—input*提供的文件，根据*—mregex*计算每个细胞中线粒体表达量，然后根据*—mini*，*--maxi*，*--mt.percent* 过滤细胞。根据*—transform*计算每个基因在每个细胞中的表达量，提取*—nf*个高变基因（在不同细胞中表达量差异较大的基因），根据高变基因在不同细胞中的表达量对这些基因进行线性降维（PCA），选取前*—dusage*个认为有生物学意义的维度进行分析。根据细胞在*—dusage*维度空间（PCs）的位置，对于每个细胞，找到距离最近的k个细胞（KNN, k nearest neighbor , k默认是50）。对于任何两两细胞,计算它们KNNs的share情况(jaccard index, 交集/并集), 作为两两细胞的相似性 (SNN, shared nearest neighbor)，然后通过leiden或louvain算法根据细胞的SNN进行聚类，不同的cluster即认为是不同的细胞类型。在这里，我们使用SNN而非直接距离的原因是单细胞转录组测序的数据稀疏性和高随机性，每个细胞只有数千个基因会被测到。

同时,为了图形展示以及交互验证，我们同时取前*—dusage*个认为有生物学意义的维度进行非线性降维(T-SNE和UMAP), 如果降维结果和聚类结果一致，则是合理的分群。原则上，我们可以在T-SNE或UMAP空间上计算KNN和SNN, 然后聚类。 但为了交互验证, 我们基本不这样做。

DoubletFinder会随机的将细胞合并成人工双胞，然后按上述流程进行分析，然后根据原始细胞与人工细胞的距离，过滤掉最可能的*—percentage*原始细胞。

**7．批次效应处理和数据整合。**

任何转录组数据均有严重的批次效应，单细胞转录组也不例外。因此，不同文库的单细胞数据必须进行批次效应去除才能都整合成一个图谱。根据2021年12月发表在《Nature methods》的《Benchmarking atlas-level data integration in single-cell genomics》，目前主流整合方法和评估如下：

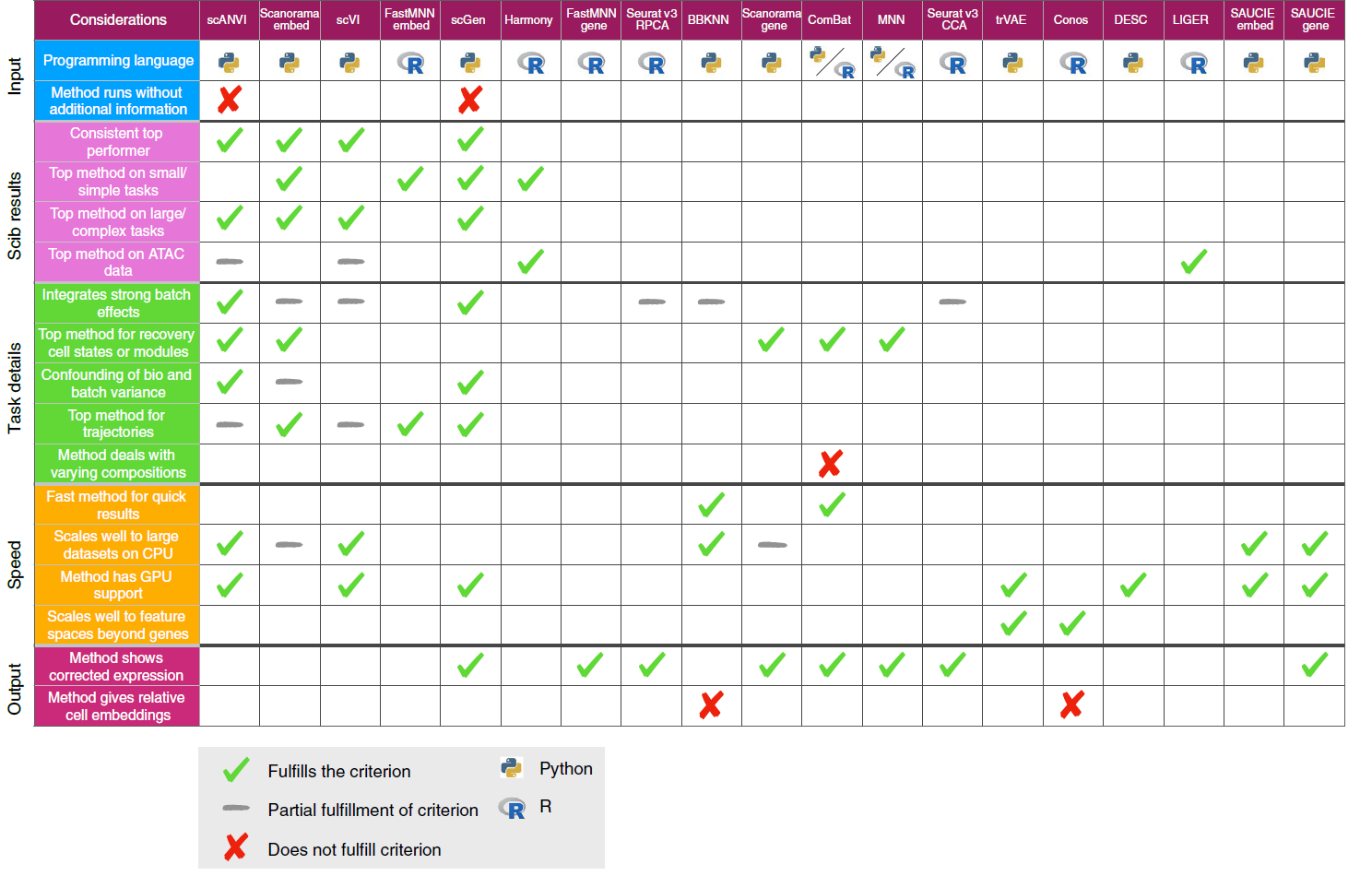


Figure 7. 主流整合方法和评价。

目前组内的提供了Seurat v3, Harmony 和 Scanorama三种整合方法。输入都是library\_name + 质控rds结果路径的列表，以及各自相应的参数。如对流程有疑问，[联系zhangpei@genomics.cn](mailto:联系zhangpei@genomics.cn)。

Seurat v3: /ldfssz1/ST\_DIVERSITY/P18Z10200N0107/zhangpei/0.bin/1.scRNA\_seurat\_pipeline/2.seurat.integration\_cca.R 效率太低，不推荐。

Harmony: /ldfssz1/ST\_DIVERSITY/P18Z10200N0107/zhangpei/0.bin/1.scRNA\_seurat\_pipeline/3.seurat\_harmony.R 适合简单整合任务，有过拟合的风险。请仔细选择高变基因数和使用的维度数（-Y）。使用的维度数过大，会导致细胞仅仅根据少数几个基因的表达量聚在一起。

Scanorama:

/ldfssz1/ST\_DIVERSITY/P18Z10200N0107/zhangpei/0.bin/1.scRNA\_seurat\_pipeline/5.scanorama.integration.R 适合复杂的任务，但需要根据样品情况和自己的判断选择合适的高变基因数，其校正后的维度一般都是有生物学意义的维度，最好全部使用。