M20高级分析调研报告

By: 张春媛

目录

[一、 背景介绍 2](#_Toc138668182)

[M20技术简介 2](#_Toc138668183)

[M20数据特征 2](#_Toc138668184)

[二、 详细内容 3](#_Toc138668185)

[可变剪切 3](#_Toc138668186)

[lncRNA 5](#_Toc138668187)

[Snp calling 8](#_Toc138668188)

[基因融合 8](#_Toc138668189)

[三、 调研结论 9](#_Toc138668190)

[四、 参考文献 11](#_Toc138668191)

**摘要**

本次调研目的是为M20数据后续高级分析提供研发方向，主要调研方向为单细胞水平的可变剪切、基因融合、snp calling以及lncRNA。本报告内容包含几个调研方向目前在单细胞水平的应用情况、M20数据的基本特征，以及相应分析针对M20数据的可行性分析等，并结合公司目前产品给出M20以外其他产品的一些建议。

总体而言，能够在单细胞水平实现的分析在单细胞水平进行测试，当理论上无法在单细胞水平实现时，可考虑在cluster水平、样本水平、组水平进行分析测试。由于M20数据质控后，平均到每个细胞中的reads数据较低，结果可信度偏低，同时考虑技术难度，建议在cluster水平进行分析；lncRNA可以单独或者和mRNA一起实现常规分析，并在展示结果的时候重点展示lncRNA的分析结果，并在后续增加共表达网络分析，结合scATAC-seq预测TF等；M20数据和10X数据均可在单细胞水平call SNP，并用于后续分析；STAR-fusion工具能够使用M20数据和10X数据在单细胞水平检测基因融合，同样可信度偏度。详细内容参考第二章，具体研发建议参考第三章。

# 背景介绍

### M20技术简介

目前单细胞RNA测序技术按照获得单细胞的方式可以分为两种，分别为droplet和plate两种类型，前者通量高，典型的技术代表为10X单细胞测序，后者每个细胞获得的数据量大，典型的技术代表为smart-seq[1-4]单细胞测序。两种平台产生的scRNA-seq数据的重要区别之一是前者只有一端reads含有cDNA信息，而后者双端reads都含有cDNA信息，这使得两个平台产出的数据在后续分析技术上有很大区别。目前公司的单细胞产品主要依托于10X单细胞测序技术，但是该技术不能很好的处理石蜡包裹（FFPE）的样本，M20技术可以从FFPE中获取单细胞（细胞核），并通过使用随机引物（Figure 1）的方式从单个细胞核中获取cDNA信息。得益于随机引物，M20技术测到的reads并不像10X测序技术那样，局限在3’端或者5’端，可以落到RNA的任意位置，在理想情况下能够覆盖全RNA（Figure 1），包括mRNA、lncRNA等。同时M20技术也利用droplet平台，实现高通量检测FFPE样本细胞。本次调研目标主要集中在单细胞水平的可变剪切分析、lncRNA进一步分析、SNP calling，以及基因融合四个方向，为M20技术在后续的研发方向提供参考。

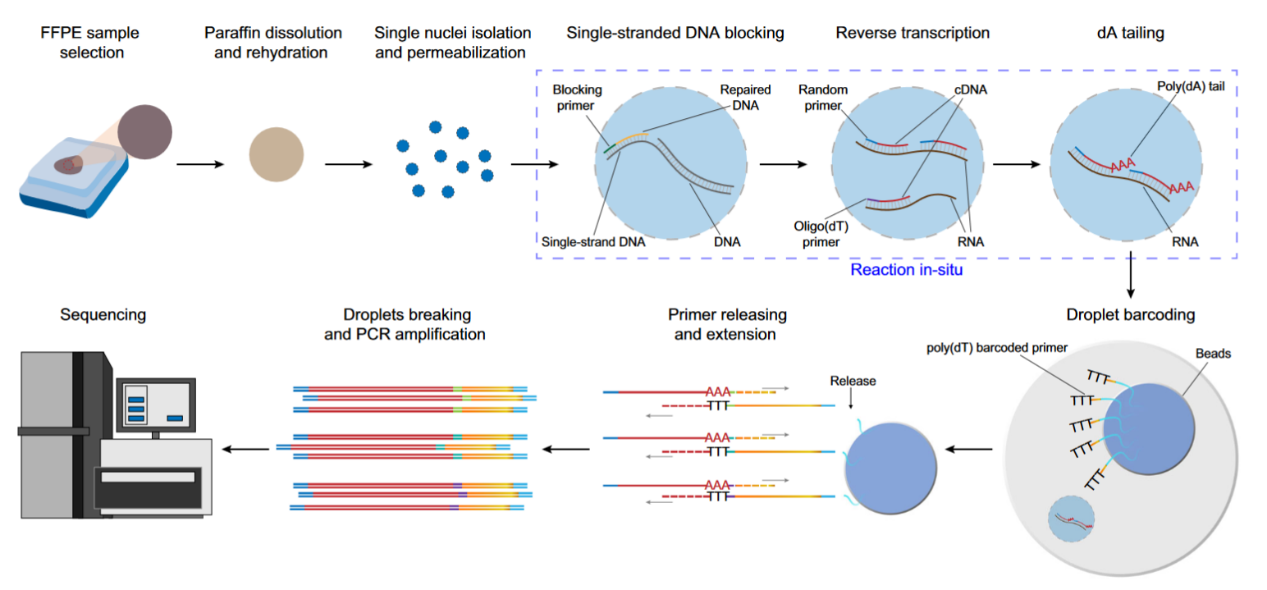


Figure ．M20技术原理。

### M20数据特征

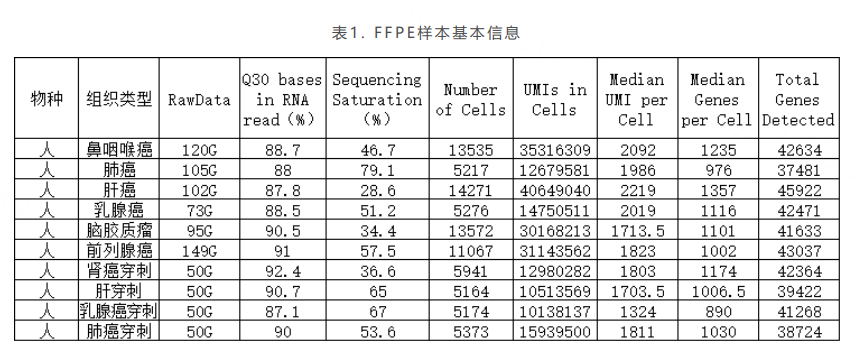
M20技术目前积累了少量数据（Table 1），结合已经公布的数据以及同M20公司员工的沟通情况，总结M20数据的特征如下：

* Read1中包含的信息有adapter、barcode、UMI，在质控后只剩余barcode和UMI信息，Read2为cDNA信息，即为单端测序信息；
* 平均每个细胞中包含reads数为2000 ~ 7000；平均到每个细胞的UMI数量为2K（10X数据平均到每个细胞的UMI数量为3K~9K）；
* RNA从细胞核中提取，比对到intron区的reads占总量的51%；
* 短读长，质控后平均长度69bp；
* 约有20%的reads可检测到junction信号；
* Reads分布可覆盖RNA全长；
* lncRNA占数据量的10%~15%；

鉴于M20的数据特征，部分分析明确无法实现：

* 每个细胞中含有的reads数量少，且短读长，无法在单细胞层面实现组装，即无法检测新的转录本或lncRNA，只能在有参的情况下检测已知的转录本或lncRNA；
* 位于内含子区域的reads占总量一半以上，这意味着大部分mRNA为不成熟，检测到的可变剪切假阳性率可能会占到一半以上 ；
* 只有一端数据含有cDNA信息，无法实现需要双端数据的分析；

Table ．M20数据统计



# 详细内容

### 可变剪切

可变剪切是单细胞分析中除表达以外大家最关心的内容之一[5]，部分基因并没有差异表达，但是却表达不同的转录本（Figure 2），甚至高达55%的同类型细胞表达不同的转录本[6]。与表达量分析不同，转录本的检测需要测序reads能够覆盖RNA全长，10X数据基于droplet平台，reads集中在3’端，并不适用于可变剪切的分析，目前大多数单细胞水平的可变剪切分析都使用plate平台的测序数据[6-12]。得益于随机引物的设计，M20数据也能够覆盖RNA全长，但是reads的数量是常规plate数据的1/100，同时仅一端reads含有cDNA信息，使得M20数据在单细胞水平做可变剪切变得十分困难。

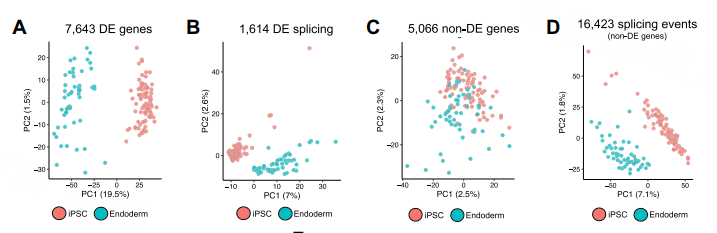


Figure ．基因表达和可变剪切对细胞分类的影响[9]。

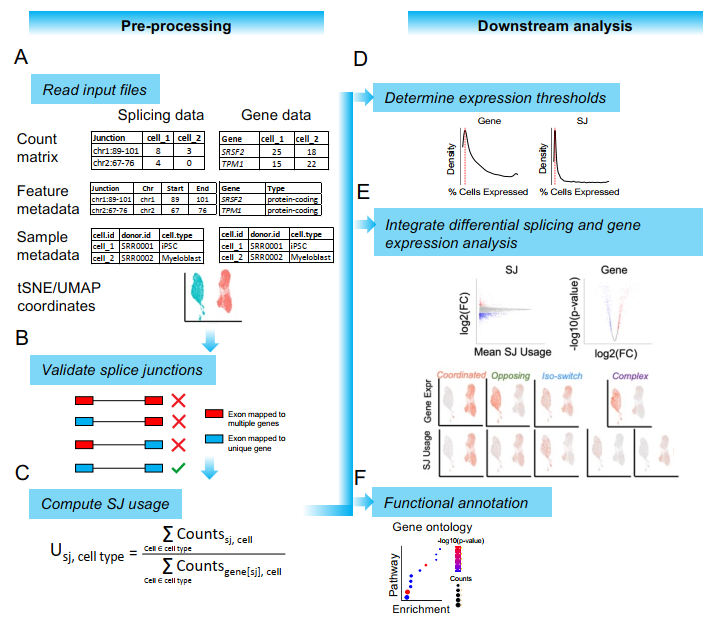


Figure ．MARVEL针对droplet平台数据做可变剪切分析流程[9]。

在以上前提下，我们同时考虑软件设计原理和检测可变剪切类型寻找可能对M20数据做可变剪切分析的工具。mRNA的可变剪切主要有7种类型，其中保留内含子（retained-intron，RI）的可变剪切类型对M20数据在可变剪切分析上的影响最大，由于M20技术RNA从细胞核中提取，大部分mRNA并不成熟，示例数据显示约有51%的reads比对到内含子区域，这会导致检测到的可变剪切大多数都是假信号，因此以MARVEL[9], VALERIE[12]为代表的对可变剪切类型做出区分的工具不能应用在M20数据在单细胞水平的分析中，仅检测skipped-exons (SE) 类型可变剪切的工具如BRIE/BRIE2[7, 8], Expedition[6], SCATS[11] 并不受内含子保留的影响，但是在当前数据量低的前提下，可变剪切分析结果的可信度并不高。DESJ-detection[10]工具并不考虑可变剪切复杂的情况，直接以检测junctions差异表达的方式检测可变剪切，但是在单细胞数据稀疏的前提下，差异表达的junction并一定是可变剪切，这可能会导致假阳性偏高。以上内容针对M20数据特征在分析意义上进行讨论，总体认为在单细胞水平做可变剪切可行性不高，此外技术门槛偏高，现有分析多针对plate平台数据进行设计，M20数据为droplet平台数据，两个平台分析流程并不通用，需要对流程中的多款软件进行调试，以及数据格式的重新整理。

虽然在单细胞水平做可变剪切分析较为困难，但是MARVEL工具专门为droplet平台数据设计了cluster水平的可变剪切分析流程（Figure 3），可同时适用于M20数据和10X数据。

**可变剪切分析建议：**

不建议在单细胞水平做可变剪切分析研发，若一定要研发，推荐测试Expedition工具。建议使用MARVEL工具在cluster水平做可变剪切分析。

### lncRNA

lncRNA长度较长，通常大于200nt，具有类似mRNA的结构，有些具有poly(A) 尾巴，有些没有poly(A) 尾巴，分化过程中有动态的表达与不同的剪接方式，与mRNA相比，lncRNA表达量更低，10X单细胞测序技术通过富集poly(A) 尾巴的方式提取RNA，使得lncRNA数据量相对较少，约占总数据量的8%，前期的分析中很少涉及到lncRNA。M20测序技术采用随机引物的方式提取RNA，更多的lncRNA被富集出来，在示例数据中，lncRNA大约占总RNA的10%~15%（Figure 4），在数据量上为后续的联合分析提供了可能。

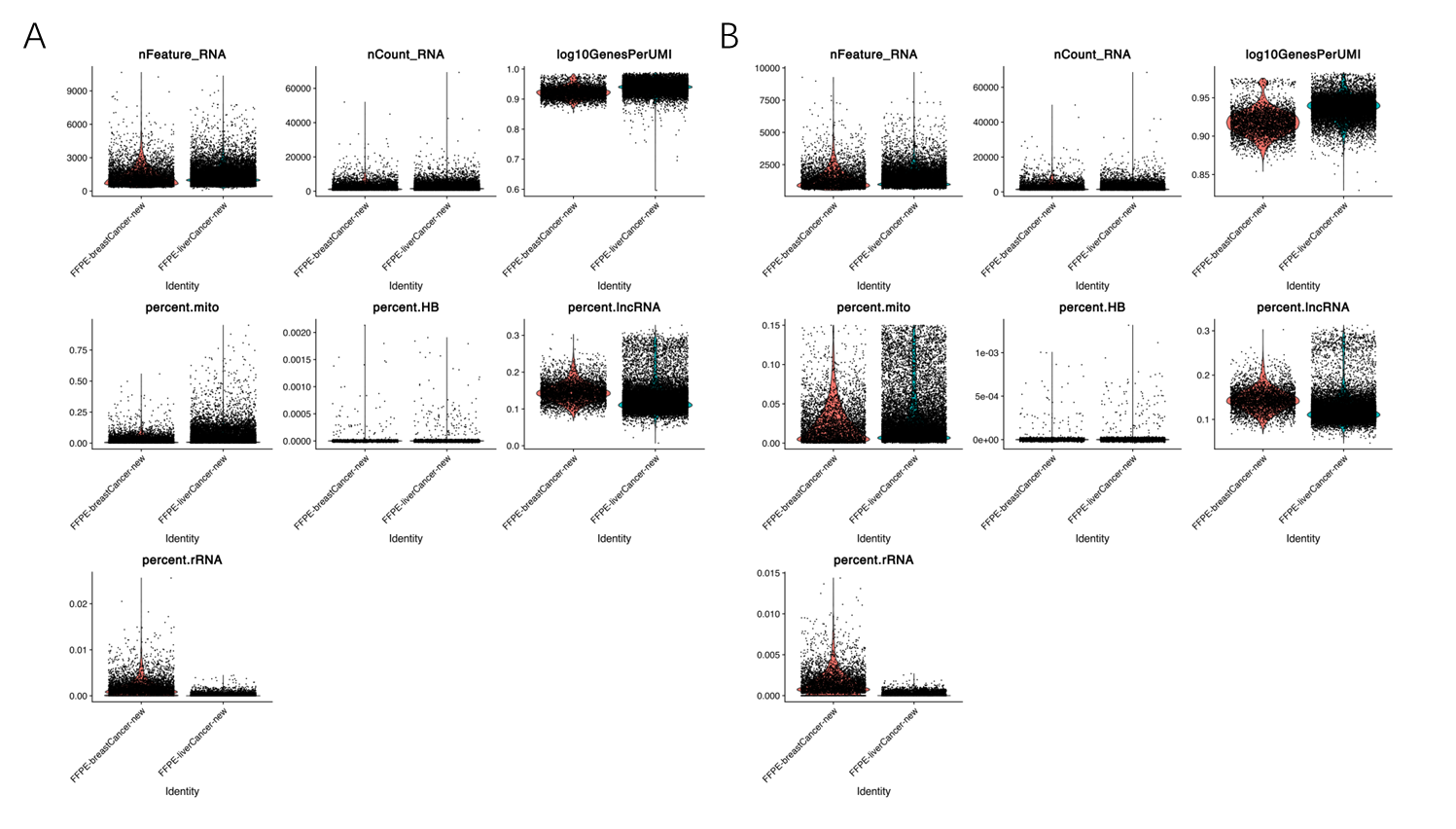


Figure ．以示例数据（乳腺癌样本和肝癌样本为例）各RNA在M20数据中的比例。（A）质控前数据分布情况；（B）质控后数据分布情况。

虽然人类的遗传数据是目前研究最为深刻和丰富的数据，但是针对lncRNA的研究还远远不足，Zhijie Wu在单细胞水平重头组装了31K条 lncRNA，但是其中约75%的lncRNA是未知的[13]，提示lncRNA的从头组装十分重要，但M20数据量太低，平均每个细胞中reads量仅有7K，因此无法在细胞水平实现组装，因此只能通过参考数据库鉴定已知的lncRNA，又因M20技术检测的是细胞核中的RNA，存在lncRNA不成熟问题，即使在cluster水平做组装，也会存在较多假阳性，因此可行性不高。

lncRNA的表达模式呈现细胞类型特异性或者组织特异性，即使在bulk RNA中低表达，也有可能在特定类型的细胞中高表达[14]，lncRNA表达量比mRNA具有更大的方差（Figure 6），虽然表达总量比mRNA小，但是其可能具有更高的差异倍数（Figure 6）[13]，因此lncRNA的差异分析可以作为后续研发的重点。

lncRNA的功能十分复杂，Becker 的文章中总结了部分真核生物中lncRNA的生物学功能[15]，总之，多是以调控的方式参与到生命活动中，共表达网络是分析功能调控的主要方式之一。Xue Zhao的综述[16]总结了单细胞多组学在网络分析中的应用，目前主要有单组学和多组学联合分析两种策略分析基因调控网络，前者使用scRNA-seq数据构建基因调控网络，并使用scATAC-seq数据做验证，后者直接将两部分数据的矩阵联合在一起进行聚类分析，并构建基因调控网络和TF相关分析。公司目前也有scATAC-seq相关产品，因此使用M20数据单独做网络分析，或者结合scATAC-seq产品做联合分析可以作为后续研发的重点。

Raza Ur Rahman另辟蹊径，开发的 Singletrome工具重新整合GENCODE数据库和LncExpDB数据库中的人类的lncRNA注释信息，同时包含已验证和未验证的lncRNA，以增加检测到的lncRNA的比例[17]，若客户的目标基因是lncRNA且常规分析手段并没有检测到目标基因时，可以使用该工具进一步尝试。

**单细胞水平lncRNA分析建议：**

将lncRNA和mRNA数据合并做常规分析，可在展示结果的时候将lncRNA单独展示。lncRNA单独进行常规分析效果未知，需要进一步测试。聚类、差异分析等目前公司有成熟的常规流程，暂时没有更换工具的必要，本次调研不再对常规分析工具做调研。网络分析也有现成的代码，不再进行赘述。在老师以lncRNA为主要分析目标时，可尝试Singletrome工具。Figure 6展示两个散点图可以增加到M20常规分析中，对在bulk水平低表达，但是在部分细胞中高表达的基因进行可视化。

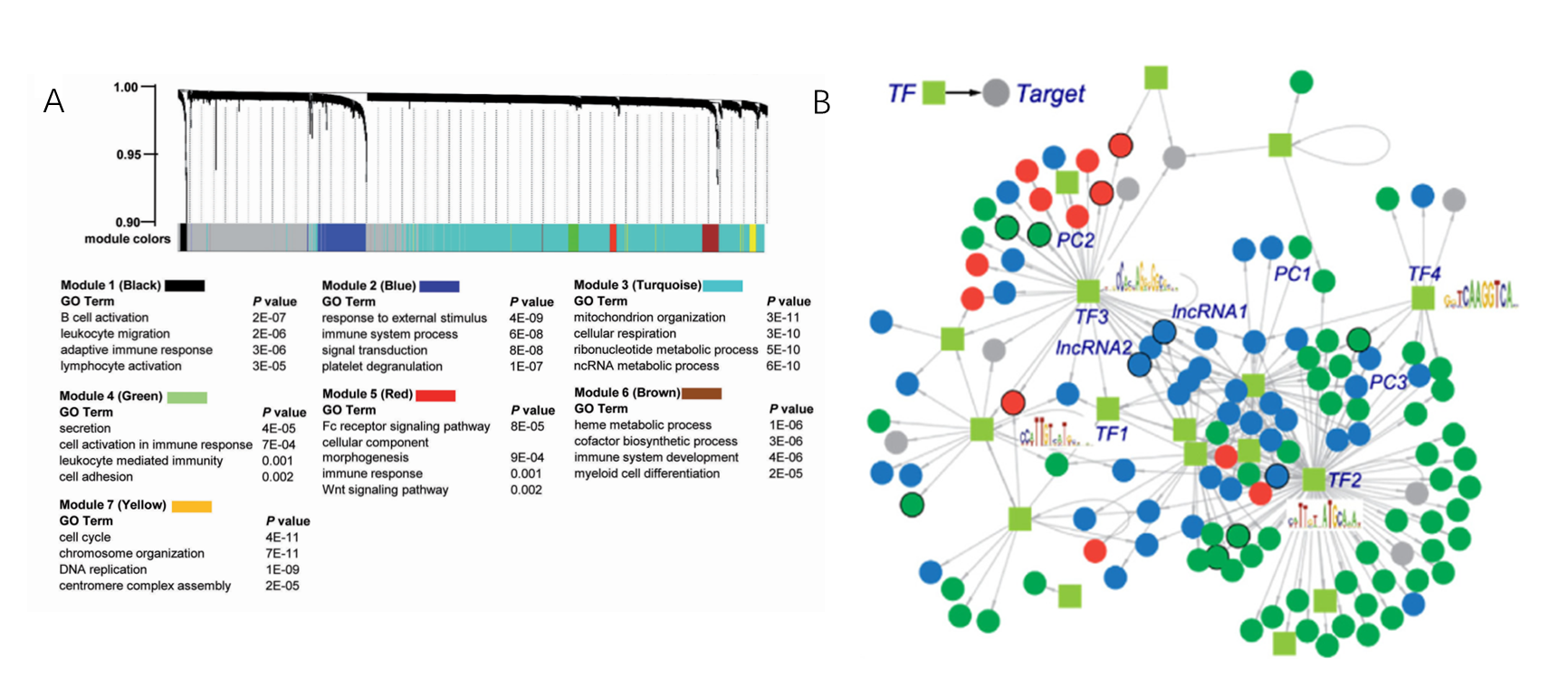


Figure ．网络分析结果展示。（A）树的方式展示模块[13]；（B）网络的方式展示模块[16]。

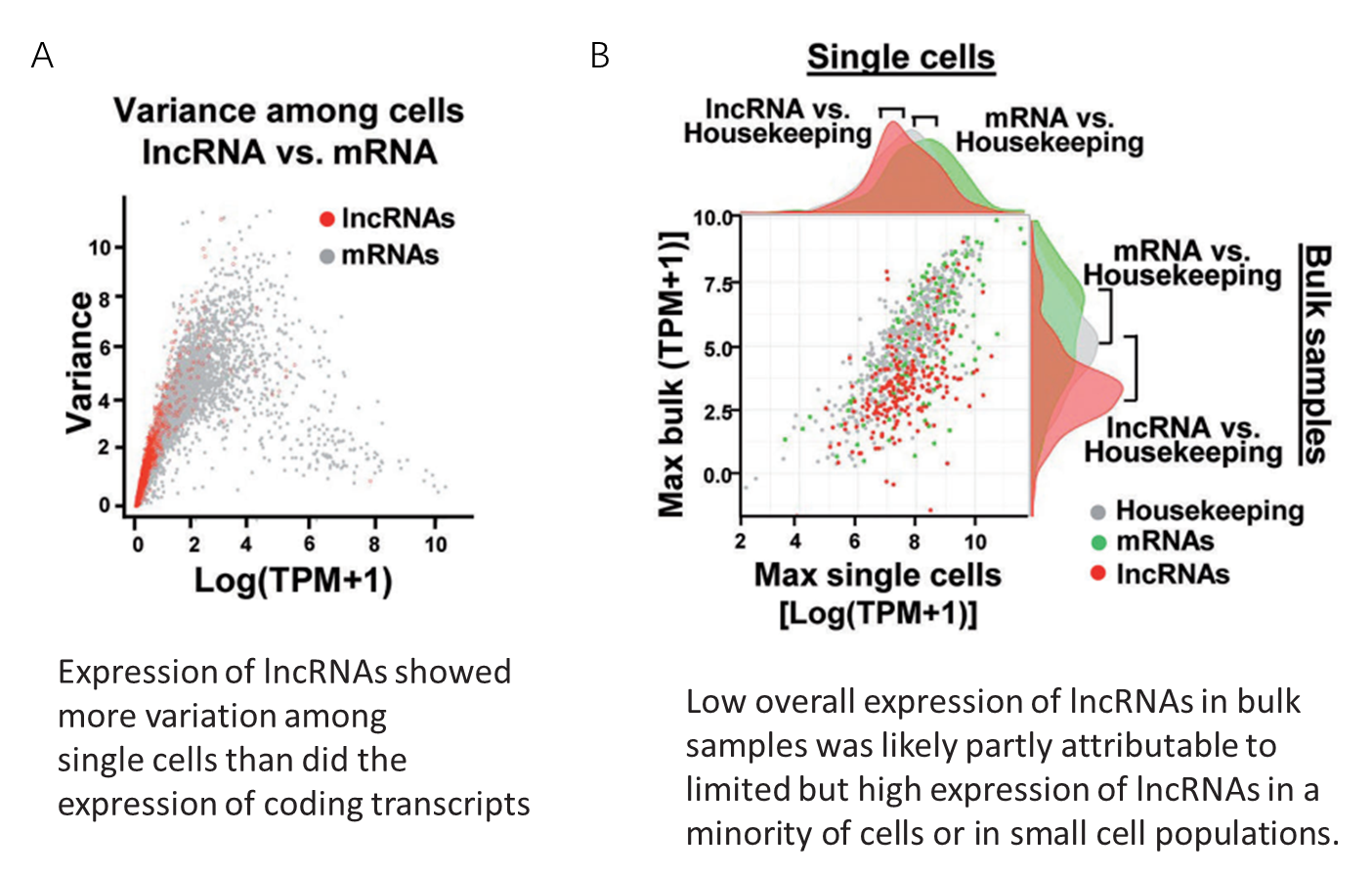


Figure ．lncRNA分析示例[13]。（A）lncRNA和mRNA表达量方差散点图，体现lncRNA更低的表达量和更高的异质性；（B）scRNA-seq和bulk RNA联合分析时最大表达量散点图，体现RNA在少量细胞中高表达。

### Snp calling

Fenglin Liu在19年发表的文章中同时使用smart-seq数据和10X数据系统的比较了当时应用较多的两款比对工具和七款SNP检测工具在检测SNP的效力[18]。M20数据与10X 数据都来自droplet平台，因此我们参考利用10X数据比较的结果挑选后续分析使用的软件。两款比对工具中，STAR的真阳性率显著高于GSNAP（Figure 7A）。七款SNP检测工具包含SAMtools, the GATK Best Practice, CTAT, FreeBayes, MuTect2, Strelka2, 以及 VarScan2，考虑到M20数据最大的短板为测序深度较低，因此对测序深度有较高要求的SAMTools, MuTect2, Strelka2, 以及 VarScan2都被排除，仅剩余FreeBayes, CTAT和GATK（Figure 7B），但是由于FreeBayes在低深度的时候假阳性较高（Figure 7C），因此也被排除。CTAT是基于GATK Best Practice设计的，Figure 7B和C显示CTAT和GATK在真阳性率和假阳性率的曲线几乎完全重叠，但是在使用上CTAT要比GATK方便的多，因此我们将CTAT作为snp calling重要的候选工具之一。Cellsnp-lite[19]专为使用单细胞RNA测序数据检测SNP而设计，可以提供常见SNP的VCF文件作为参考快速检测常见变异，也可以自行检测SNP，只是该工具与SAMtools同源，在droplet的数据分析中，当测序深度较低时SAMtools的真阳性率不足，目前尚缺少对该软件评估。近期发表的工具scAllele[20]不仅可以检测SNP，还可以检测小插入与缺失、等位基因特异的可变剪切分析，但是该方法目前并不适用于droplet的scRNA-seq数据。

**单细胞水平snp calling建议**：

M20数据在单细胞水平上的SNP calling候选工具暂定STAR/STARsolo + CTAT - Mutations Pipeline。

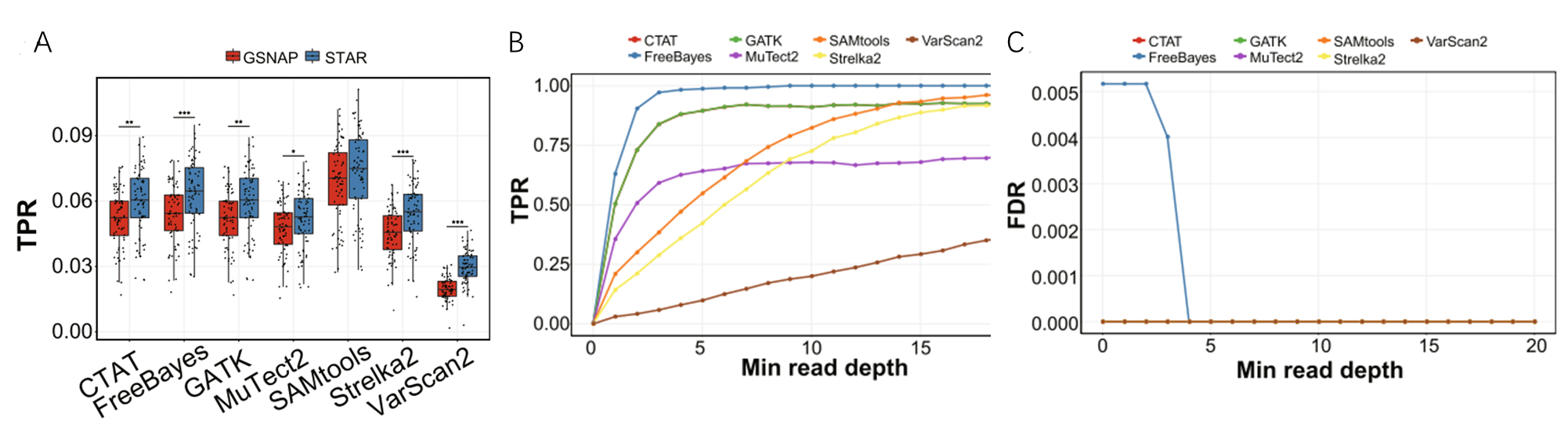


Figure ．Fenglin Liu使用10X 单细胞测序数据检测SNP工具比较 [18]。

### 基因融合

基因融合在肿瘤形成和发展过程中可能扮演重要角色。基因融合分析不仅也面临着数据量低天然引入的不精确性，还面临另外一个技术性问题，基因融合检测一般会同时用到split reads和discordant reads（双端），鉴于M20数据只有一端reads含有cDNA信息，完全缺失discordant reads，因此专为双端数据设计的工具，如scFusion[21]、FusionCatcher[22]、EricScript[23]、Arriba[24]等便不能应用于M20数据。FusionMap[25]虽然兼容单端测序数据的工具，但并不适用于droplet平台产出的数据。STAR-Fusion[26, 27] 从v1.5.0版本开始能够适用于droplet平台产出数据，可同时应用于M20数据和10X数据。

**基因融合分析建议：**

可尝试在单细胞水平进行基因融合分析，但是考虑到10X数据集中在3’端，M20数据量比10X还要少，且STAR-Fusion采用拼接的方式处理单端reads进而检测基因融合，可以想见结果的可信度并不高。

# 调研结论

**针对M20单组学数据分析建议：**

1. mRNA和lncRNA一起做常规分析，比较仅使用mRNA、仅使用lncRNA、同时使用mRNA和lncRNA时的聚类效果；【推荐指数★★★★★】
2. 增加lncRNA 和 mRNA表达量比较（Figure 6A）；【推荐指数★★★★★】
3. 使用mRNA和lncRNA构建调控网络，注释模块功能，预测转录因子；【推荐指数★★★★★】
4. 在单细胞水平call SNP，候选工具为STAR/STARsolo + CTAT - Mutations Pipeline（适用于10X）； 【推荐指数★★★★★】
5. 使用MARVEL工具在cluster水平做可变剪切分析（适用于10X）【推荐指数★★★★★】
6. 在单细胞水平水平检测基因融合，候选工具有STAR-Fusion 【推荐指数★★】
7. 测试Expedition工具在单细胞水平检测可变剪切，由于M20为单端数据，需要重新调整比对策略，因此测试难度较高；【推荐指数★】
8. 测试DESJ-detection工具在单细胞水平检测可变剪切，难度同上；【推荐指数★】
9. 测试Singletrome，查看lncRNA增加效果；【推荐指数★】

**M20数据和其他组学数据联合分析建议**：

1. bulk RNA和scRNA-seq联合分析增加基因最大表达量在两个组学间的散点图（Figure 6B）；【推荐指数★★★★★】
2. 差异表达基因随着细胞发育表达量变化热图（Figure 8）；【推荐指数★★★】

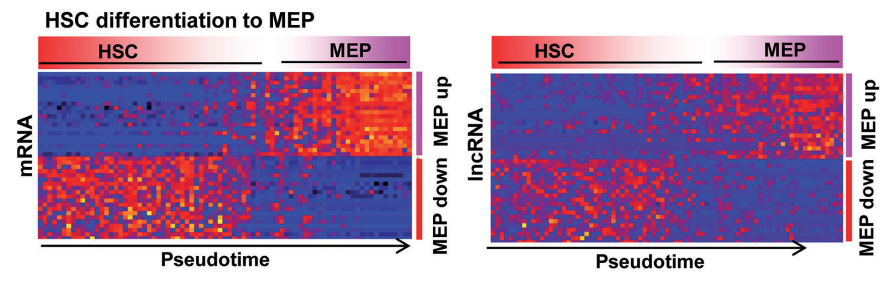


Figure ．差异基因表达随着细胞分化的变化

参考链接：

Is it possible to call SNPs or splicing variants from single cell gene expression data using 10x software?

<https://kb.10xgenomics.com/hc/en-us/articles/360016405671-Is-it-possible-to-call-SNPs-or-splicing-variants-from-single-cell-gene-expression-data-using-10x-software->

cellsnp-lite model1 已经处理了人类常见的snp参考VCF文件，可以直接下载使用

<https://cellsnp-lite.readthedocs.io/en/latest/snp_list.html>

# 参考文献

1. Goetz, J.J. and J.M. Trimarchi, *Transcriptome sequencing of single cells with Smart-Seq.* Nature Biotechnology, 2012. **30**(8): p. 763-765.

2. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nature Protocols, 2014. **9**(1): p. 171-181.

3. Hagemann-Jensen, M., C. Ziegenhain, and R. Sandberg, *Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress.* Nature Biotechnology, 2022. **40**(10): p. 1452-1457.

4. Hagemann-Jensen, M., et al., *Single-cell RNA counting at allele and isoform resolution using Smart-seq3.* Nature Biotechnology, 2020. **38**(6): p. 708-714.

5. Stark, R., M. Grzelak, and J. Hadfield, *RNA sequencing: the teenage years.* Nature Reviews Genetics, 2019. **20**(11): p. 631-656.

6. Song, Y., et al., *Single-Cell Alternative Splicing Analysis with Expedition Reveals Splicing Dynamics during Neuron Differentiation.* Mol Cell, 2017. **67**(1): p. 148-161 e5.

7. Huang, Y. and G. Sanguinetti, *BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments.* Genome Biology, 2021. **22**(1): p. 251.

8. Huang, Y. and G. Sanguinetti, *BRIE: transcriptome-wide splicing quantification in single cells.* Genome Biology, 2017. **18**(1): p. 123.

9. Wen, Wei X., A.J. Mead, and S. Thongjuea, *MARVEL: an integrated alternative splicing analysis platform for single-cell RNA sequencing data.* Nucleic Acids Research, 2023. **51**(5): p. e29-e29.

10. Liu, S., et al., *Single-cell differential splicing analysis reveals high heterogeneity of liver tumor-infiltrating T cells.* Scientific Reports, 2021. **11**(1): p. 5325.

11. Hu, Y., K. Wang, and M. Li, *Detecting differential alternative splicing events in scRNA-seq with or without Unique Molecular Identifiers.* PLOS Computational Biology, 2020. **16**(6): p. e1007925.

12. Wen, W.X., A.J. Mead, and S. Thongjuea, *VALERIE: Visual-based inspection of alternative splicing events at single-cell resolution.* PLOS Computational Biology, 2020. **16**(9): p. e1008195.

13. Wu, Z., et al., *Long noncoding RNAs of single hematopoietic stem and progenitor cells in healthy and dysplastic human bone marrow.* Haematologica, 2019. **104**(5): p. 894-906.

14. Liu, S.J., et al., *Single-cell analysis of long non-coding RNAs in the developing human neocortex.* Genome Biology, 2016. **17**(1): p. 67.

15. Becker, J., et al., *What has single-cell transcriptomics taught us about long non-coding RNAs in the ventricular-subventricular zone?* Stem Cell Reports, 2023. **18**(1): p. 354-376.

16. Zhao, X., Y. Lan, and D. Chen, *Exploring long non-coding RNA networks from single cell omics data.* Comput Struct Biotechnol J, 2022. **20**: p. 4381-4389.

17. Raza Ur, R., et al., *Singletrome: A method to analyze and enhance the transcriptome with long noncoding RNAs for single cell analysis.* bioRxiv, 2022: p. 2022.10.31.514182.

18. Liu, F., et al., *Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data.* Genome Biology, 2019. **20**(1): p. 242.

19. Huang, X. and Y. Huang, *Cellsnp-lite: an efficient tool for genotyping single cells.* Bioinformatics, 2021. **37**(23): p. 4569-4571.

20. Quinones-Valdez, G., et al., *scAllele: A versatile tool for the detection and analysis of variants in scRNA-seq.* Science Advances. **8**(35): p. eabn6398.

21. Jin, Z., et al., *Single-cell gene fusion detection by scFusion.* Nature Communications, 2022. **13**(1): p. 1084.

22. Nicorici, D., et al., *FusionCatcher – a tool for finding somatic fusion genes in paired-end RNA-sequencing data.* bioRxiv, 2014: p. 011650.

23. Benelli, M., et al., *Discovering chimeric transcripts in paired-end RNA-seq data by using EricScript.* Bioinformatics, 2012. **28**(24): p. 3232-9.

24. Uhrig, S., et al., *Accurate and efficient detection of gene fusions from RNA sequencing data.* Genome Res, 2021. **31**(3): p. 448-460.

25. Ge, H., et al., *FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution.* Bioinformatics, 2011. **27**(14): p. 1922-1928.

26. Brian, J.H., et al., *STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq.* bioRxiv, 2017: p. 120295.

27. Haas, B.J., et al., *Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods.* Genome Biology, 2019. **20**(1): p. 213.