单细胞高级分析研发记录

By：张春媛

目录

[【Snp calling研发】 3](#_Toc149640686)

[【可变剪切】 3](#_Toc149640687)

[**问题1 ：为什么要做单细胞水平的可变剪切？单细胞分析最大的优势就是有细胞类型。【已完成】** 4](#_Toc149640688)

[**问题2 整理MARVEL官方pipeline，整理分析流程。【已完成】** 4](#_Toc149640689)

[**问题3 ：在现有pipeline的基础上我们需要进一步考虑的内容有哪些？** 5](#_Toc149640690)

[**问题4 ：相较原有pipeline，当前分析流程做了哪些修改？** 6](#_Toc149640691)

[**问题5 ：可变剪切分析流程数据流示意图** 6](#_Toc149640692)

[**问题6** ：**M20分析流程比对的bam对于可变剪切分析缺少tag 【已解决】** 7](#_Toc149640693)

[**问题7 ：M20数据从比对bam中提取的数据量和预期差别过大，为什么？【已解决】** 8](#_Toc149640694)

[**问题8** ：**探究cellranger和STAR的关系，cellranger是如何整合工具的** 9](#_Toc149640695)

[**问题9 ：用实际10X项目做测试，遇到STAR版本和参考版本不同不能运行的问题。【已解决】** 9](#_Toc149640696)

[**问题10 ：使用STAR（2.7.10b）提取gene和SJ时，需要指定barcode whitelist，但是whitelist在文库版本之间3’5’之间存在区别，如何保证我们使用的whitelist是正确的？【已解决】** 10](#_Toc149640697)

[**问题11 ：在提交STAR提取gene和SJ时，总是报内存不足【已解决】** 11](#_Toc149640698)

[**问题12 ：使用MARVEL pipeline 使用的原始数据做分析，摸索pipeline略过的分析部分，以对比结果的方式判断分析是否有问题**。 12](#_Toc149640699)

[**问题13 ：从sra直接提取的fastq双端reads都是150bp，数据无需要进一步进行trim【已解决】** 12](#_Toc149640700)

[**问题14** ：**比对时每个样本分别比对，后续多个样本是如何整合在一起的**？【**已解决**】 13](#_Toc149640701)

[**问题15** ：**标准化和批次效应的关系**？**差异可变剪切的分析是否需要矫正批次效应？**【**已解决**】 13](#_Toc149640702)

[**问题16** ：**目前10X和M20的数据Feature文件中gene不唯一，明确处理方案** 14](#_Toc149640703)

[**问题17** ：**从seurat对象转到MARVEL对象？【已解决】** 14](#_Toc149640704)

[**问题18** ：**gtf文件格式问题：STAR（2.7.10b）使用的gtf文件和cellranger中较老版本的不同，在MARVEL的R脚本中读取现有的gtf文件报错？【已解决】** 15](#_Toc149640705)

[**问题19** ：**MARVEL中注释不到基因信息？【已解决】** 15](#_Toc149640706)

[**问题20 ：M20数据SJ矩阵过大，有否办法缩减一下？【已解决】** 16](#_Toc149640707)

[**问题21 ：如何确定差异比较类型？样本间，cluster间，同一个cluster不同样本间？【已解决，可依据上游信息随意指定比较组】** 16](#_Toc149640708)

[**问题22 ：得到junction表达量、gene表达量的总表【已完成】** 17](#_Toc149640709)

[**问题23 ：绘制火山图【已完成】** 17](#_Toc149640710)

[**问题24 ：最低表达量参数设置【已完成，使用较低标准参数，给出尽可能多的结果给老师】** 17](#_Toc149640711)

[**问题25 ：差异junction的评估方式有两种：log2fc和delta，用哪个比较好？【已解决，delta更好】** 19](#_Toc149640712)

[**问题26 ：差异基因的筛选使用pvalue和qvalue? 【已解决】** 19](#_Toc149640713)

[**问题27 ：可变剪切的类型统计【已完成】** 19](#_Toc149640714)

[**问题28 ：可变剪切基因的筛选标准？【已完成】** 20](#_Toc149640715)

[**问题29 ： junction面板图型绘制逻辑修改。【已解决】** 20](#_Toc149640716)

[**问题30 ： 基因结构图绘制：一方面pipeline没有提供有多个junction时的绘图代码，另一方面绘制单个junction的源码已不能使用** 21](#_Toc149640717)

[**问题31 ：表达量面板图和基因结构图中的junction需要一一对应【已解决】** 22](#_Toc149640718)

[**问题32 ：10X 数据 SJ 缩减，降低资源消耗【已解决】** 22](#_Toc149640719)

[**问题33 ：M20数据量过大， SJ矩阵处理困难，要怎么缩减矩阵？【已解决】** 22](#_Toc149640720)

[**问题34 ：10X数据和M20数据在检测到可变剪切上有显著差异么？【搁置】** 23](#_Toc149640721)

[**问题35 ：能否用junction进一步对cluster进行降维，从而坚定细胞亚型。** 23](#_Toc149640722)

[**问题36 ：其他需要注意的问题** 23](#_Toc149640723)

本文档仅记录研发过程，包含每个阶段遇到的问题，如何解决的，数据特征，少量分析建议等。

# 【Snp calling研发】

研发调研在单细胞水平call SNP，候选工具为STAR/STARsolo + CTAT - Mutations Pipeline，benchmark文章明确表示该工具可以分析10X genomics单细胞数据，但是我们没有找到明确的pipeline，暂时搁置，使用有明确在单细胞水平call SNP 的Cellsnp-lite流程。

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

Git地址：<https://github.com/single-cell-genetics/cellsnp-lite>

参考文献：<https://academic.oup.com/bioinformatics/article/37/23/4569/6272512>

**问题**：拿到单细胞水平的SNP后还需要进一步的分析，只是单纯的VCF文件对客户而言并没有什么意义，考虑工作进度的原因，进一步分析调研和研发至少要安排在可变剪切完成之后进行。

# 【可变剪切】

难点：

1. MARVEL pipeline 从中间开始，略过了从原始数据到中间文件的过程，需要自己摸索
2. M20和10X流程使用工具不同，需要保证两个流程结果没有大的区别
3. 候选基因的筛选标准确定

麻烦点：

MARVEL pipeline十分简单，需要针对公司目前单细胞分析的产品制定新的分析内容

现有pipeline中大部分代码已经失效，不能直接获得结果

**工具**：我们选择MARVEL在cluster水平上分析可变剪切，可以同时应用于10X数据和M20数据。

**参考文献**：<https://academic.oup.com/nar/article/51/5/e29/6985826>

**Git地址**：<https://github.com/wenweixiong/MARVEL>

**官方pipeline地址**：[MARVEL (wenweixiong.github.io)](https://wenweixiong.github.io/MARVEL_Droplet.html#Gene-splicing_dynamics)

**服务器工具安装**：由于其使用R包完成分析，依赖的包有很多，因此我们专门创建一个R环境，将所有使用的工具都安装到该环境下，环境路径为：/public/dev\_scRNA/zhangchunyuan/tools/R4.2.0\_MARVEL，

激活环境前需要执行 source ~/miniconda3/etc/profile.d/conda.sh， 再conda activate ~/zhangchunyuan/tools/R4.2.0\_MARVEL/ 。未来配置docker时可以直接拷贝该环境。

1. **：为什么要做单细胞水平的可变剪切？单细胞分析最大的优势就是有细胞类型。【已完成】**

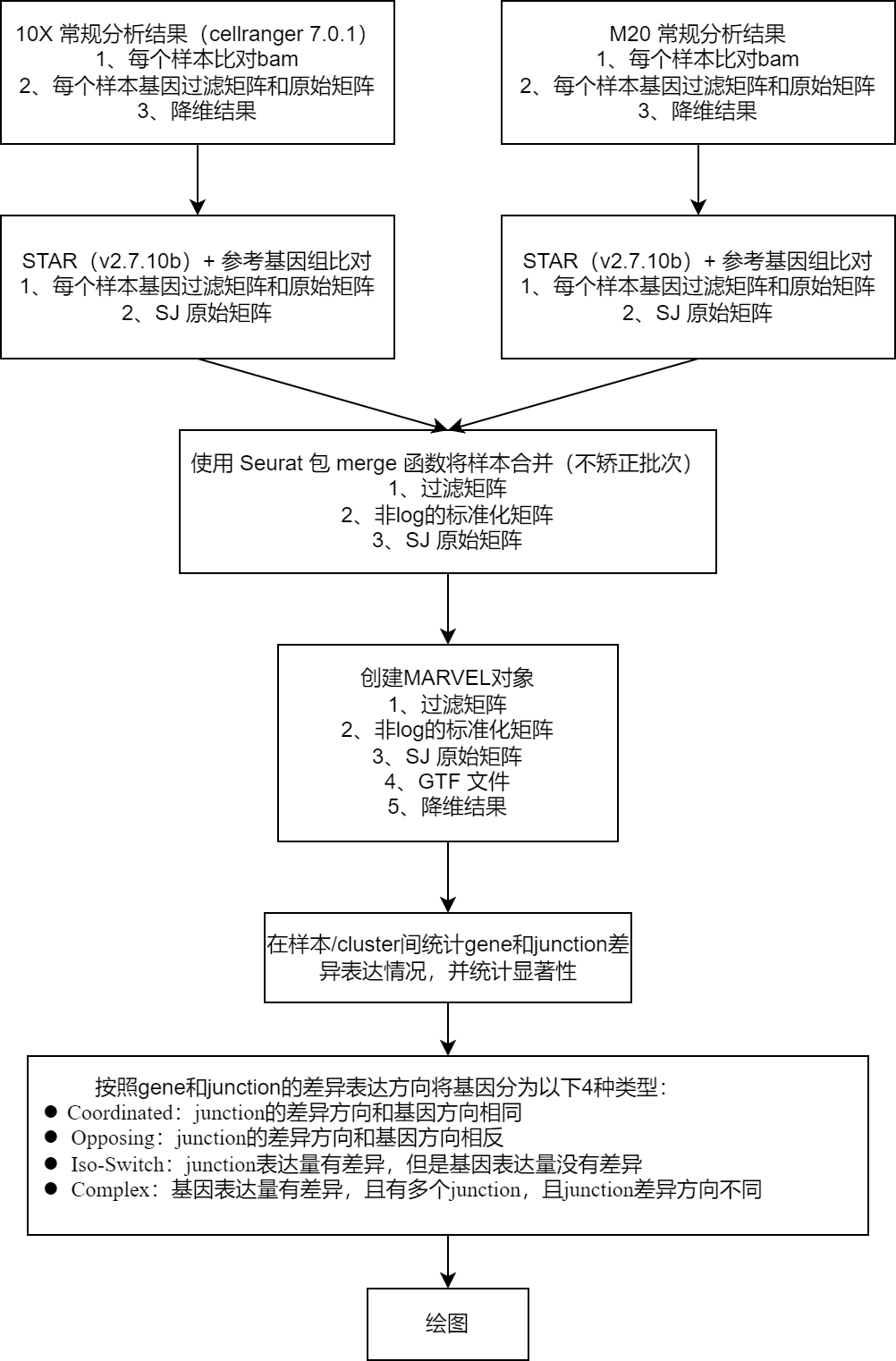
单细胞可变剪切相对于传统可变剪切的优势在于能够在细胞类型之间进行比较，如果样本包含的细胞类型不止一种，那么传统转录组将所有RNA合并在一起进行分析的方式就会抹杀掉细胞类型信息。本次开发的marvel流程，能够在样本之间，也可以在cluster之间检测可变剪切，这与目前单细胞常规差异基因分析的策略一致。

鉴于当前可变剪切尚不能检测新的细胞类型，只能在已经确定好细胞类型后检测细胞类型之间的可变剪切，当老师的样本是已经分好的细胞类型，那么marvel可变剪切的理论优势并不比传统方式更优。

1. **整理MARVEL官方pipeline，整理分析流程。【已完成】**
2. 每个样本分别Cellranger比对，STAR提取gene和SJ信息，分别存储在Solo.out/Gene/filtered/ 和 Solo.out/SJ/raw/ 文件夹下，并将4个样本的barcode、feature、matrix信息整合在一起。【没有说明如何整合，仅提供了已经整合的示例数据】
3. 使用SingCellaR工具，对已经整合后的gene数据做标准化。【not log2-transformed】
4. 后续分析共需要3个数据集：Solo.out/Gene/filtered/文件夹下没有标准化的数据集、已经标准化后的数据集、Solo.out/SJ/raw/文件夹下未做任何处理的可变剪切数据集。注意数据集之间细胞ID需要一一对应。
5. 读取降维坐标
6. 读取GTF文件，用于对可变剪切做注释
7. 将以上5个数据创建MARVEL对象，进行后续分析
8. 查看基因表达率和可变剪切表达率的密度图确定进一步分析的基因。
9. 差异表达可变剪切火山图、差异表达可变剪切类型注释、在降维图上可视化、GO注释、最高差异表达可变剪切基因的进一步展示。
10. **：在现有pipeline的基础上我们需要进一步考虑的内容有哪些？**
11. 安装所有工具
12. 10X数据可以直接使用STAR提取信息，M20数据质控过程需要摸索参数
13. 多个样本的信息是如何整合在一起的？
14. Feature文件中gene不唯一，要如何处理？
15. 10X使用Gene，M20使用GeneFull，是否会对结果数据量带来系统性影响？
16. M20分析中使用的是cellranger默认的质控方式，M20同时对细胞和基因都做了质控，要如何一一对应？
17. 示例文件在实验阶段就已经明确细胞类型了，因此在样本之间比较和细胞类型之间的比较是统一的，在我们的分析中，需要提前确认比较类型，目前可以选择的比较类型有：**样本间比较**、**样本间同一个cluster间的比较**、**同一个样本不同cluster间的比较。**
18. 需要根据样本量调整的内容有：只有一个样本的时候，只能在同一个样本不同cluster间比较可变剪切；当两个分组都有多个样本的时候是否需要考虑批次效应？若考虑，选择何种方式去批次？
19. MARVEL pipeline中通过查看基因和可变剪切表达率密度图的方式筛选参与后续分析的基因数量，但是对于公司流程不方便使用这样的方式，一方面会增加人工，另一方便不同的人使用的标准会存在差异，不便于统一，因此需要探究一个自动化的方式。
20. 进一步分析。
21. **：相较原有pipeline，当前分析流程做了哪些修改？**

* 我们的流程在绘图时以基因为单位
* 火山图增加legend，增加以log2fc为Y轴的展示方式
* 表达量向降维映射的时候也将基因和junction合并在一起进行展示
* gene和junction的表达面板图整合为一个图，分别将基因名称、junction名称作为X轴
* 结构图自动添加多个junction，并添加legend

1. **：可变剪切分析流程数据流示意图**



1. ：**M20分析流程比对的bam对于可变剪切分析缺少tag 【已解决】**

需要在M20比对时增加tag：CY UY

1. **：M20数据从比对bam中提取的数据量和预期差别过大，为什么？【已解决】**

**数据**：10X分析流程中，使用STAR从cellranger比对得到的bam文件中提取SJ，其中

|  |  |  |  |
| --- | --- | --- | --- |
| Sample | Cellranger filtered/barcodes.tsv.gz 数量 | Splicing filtered/barcodes.tsv.gz 数量 | 减少百分比 |
| SRR9008752 | 9555 | 8656 | 9% |
| SRR9008753 | 7284 | 5970 | 18% |
| SRR9008754 | 13546 | 11296 | 17% |
| SRR9008755 | 8092 | 6474 | 20% |

但是M20数据减少量过多了，从2万，减少到2千，减少的太多了，需要检查怎么回事。

**测试方案**：STAR在初次比对的时候要比我们再二次比对的时候多几个参数，分别增加测试以下几个参数对filtered/barcodes.tsv.gz数据量的多少

|  |  |  |
| --- | --- | --- |
| 新增参数 | Splicing filtered/barcodes.tsv.gz 数量 | 默认参与及新增参数意义 |
| 第一次比对【M20常规输出】 | 20079 | - |
| 原始参数 | 1899 | - |
| --soloStrand Reverse | 10301 |  |
| --soloUMIdedup Exact | 1906 | 默认UMI比对时允许一个错配，当前参数不再允许错配 |
| --outMultimapperOrder Random | 1899 |  |
| --soloStrand Reverse  --soloUMIdedup Exact | 9769 | 用时：  real 30m24.727s  user 199m4.265s  sys 11m1.007s |
| --soloStrand Reverse  --soloUMIdedup Exact  --outMultimapperOrder Random | 9769 | 用时：  real 30m30.056s  user 182m56.787s  sys 7m6.362s |

**结论**：--soloStrand Reverse参数是建库过程中是否经过逆转录，M20数据必须添加该参数；--soloUMIdedup Exact 不允许UMI错配，会使检测到的细胞数量稍微变少，需要和M20流程保持一致，因此也添加【尚不明确cellranger比对时候是否UMI也不允许错配】；--outMultimapperOrder Random 对结果本身没有任何影响，不排序能够稍稍节省一些计算资源，因此也添加上去。同时添加三个参数。

1. ：**探究cellranger和STAR的关系，cellranger是如何整合工具的**

Cellranger count merge

1. **：用实际10X项目做测试，遇到STAR版本和参考版本不同不能运行的问题。【已解决】**
2. Cellranger比对
3. STAR提取 Gene 和 SJ —— 报错

**报错原因**：cellranger内部整合了STAR工具，对于10X数据而言，如果比对时使用的cellranger的STAR版本和外面使用的SRAT版本不一致，可能会造成报错，本次报错是因为目前 scrna\_pipeline 使用cellranger版本为7.0.1，其使用的STAR版本为2.7.4a，但是我们使用的外部STAR版本为2.7.10b，主要问题是旧版本创建的参考基因组对于新版本而言是不完整的，因此解决办法为：1、使用旧版本的STAR；2、使用新版STAR重新创建一下参考文件。

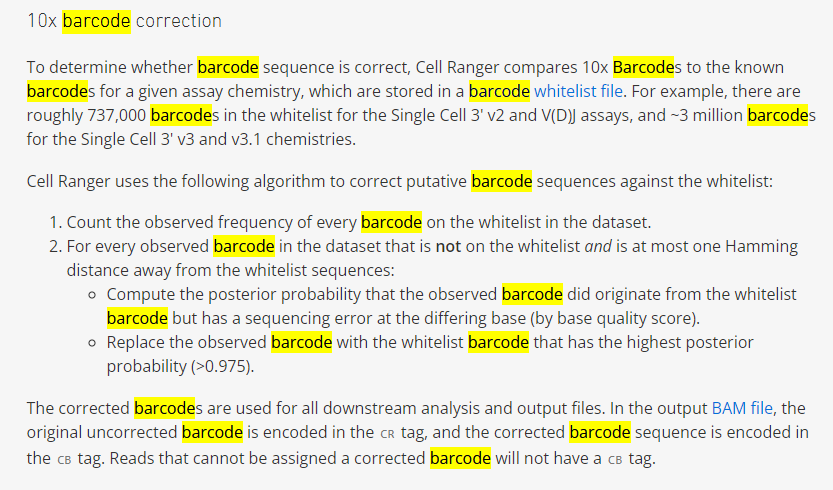
**陆瑶回复关于cellranger版本问题**：7.1和7.0在单细胞定量方面升级不是太大，实际生产中为了便于同一项目各样本的软件版本保持同一，如果没有大升级，一般不会频繁更新。（单细胞项目收样时间跨度大，有的半年到一年内会分批送样，中途换了新版本后面需要人工统一修改，不太利于大批量生产）

**孙坤回复关于参考的问题**：由于M20使用的是随机引物，因此M20数据能够覆盖到基因的更多位置上，因此在比对的时候需要对M20构建新的参考基因组，而不能使用旧的参考基因组。M20的参考基因组比cellranger包含更多的内容，cellranger的参考基因组是M20的子集。

**解决方案**：为了尽可能让M20流程和10X流程保持一致，10X数据比对后，使用STAR提取gene和SJ，使用M20使用的STAR（2.7.10b）工具和参考基因组。

1. **：使用STAR（2.7.10b）提取gene和SJ时，需要指定barcode whitelist，但是whitelist在文库版本之间3’5’之间存在区别，如何保证我们使用的whitelist是正确的？【已解决】**

**问题来源**：Cellranger工具包中含有多个barcode whitelist，自己有一套判断标准去寻找正确的barcode，参考下图：

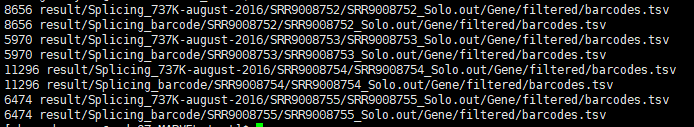


但是我们并不能实现cellranger这么负责的功能，也没有必要，那么如何选择正确的barcode whitelist呢？

**候选解决方案**：1、明确公司目前都有哪些产品，使用哪个barcode whitelist文件，用config增加参数判断；2、从bam文件中找read1对应的barcode，提取出来，制作一个barcode whitelist；3、读取文件raw\_feature\_bc\_matrix/barcodes.tsv.gz中的信息，将该文件作为barcode whitelist使用。

**优缺点**：**方案1**，不会造成任何信息丢失，但是需要明确不同文库时使用的白名单，保证不会出错，还要注意生产工具升级的时候是否有新的白名单升级；**方案2**，避免了用错白名单造成的系统性错误，但是可能会造成信息量丢失，具体信息会丢失多少，尚不明确。目前尚不知道bam文件中包含的barcode信息是否和raw\_feature\_bc\_matrix/barcodes.tsv.gz相同，也需要进一步的数据支持。

**数据支持**：



**结论**：使用正确的白名单和直接使用 raw\_feature\_bc\_matrix/barcodes.tsv.gz 输出的结果中的数据量是相同的，不存在信息丢失问题，因此不再处理白名单的问题，直接使用cellranger输出的raw\_feature\_bc\_matrix/barcodes.tsv.gz文件；同时SJ矩阵只有raw矩阵，使用raw\_feature\_bc\_matrix/barcodes.tsv.gz作为白名单也会降低SJ矩阵的大小，节省内存。

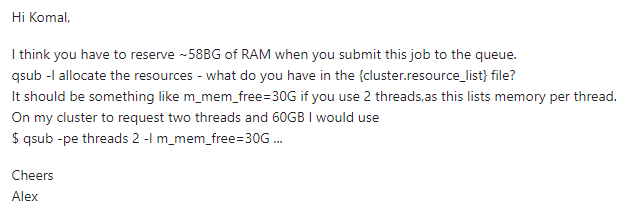
1. **：在提交STAR提取gene和SJ时，总是报内存不足【已解决】**

**报错信息**：EXITING: fatal error trying to allocate genome arrays, exception thrown: std::bad\_alloc。Possible cause 1: not enough RAM. Check if you have enough RAM 31368512329 bytes；Possible cause 2: not enough virtual memory allowed with ulimit. SOLUTION: run ulimit -v 31368512329

**解决过程**：

1、按照报错信息中的建议，在snakemake流程的shell中添加ulimit -v 31368512329，但是我们没有权限——不通；

2、将报错信息帖到搜索框中检索，找到网址<https://github.com/alexdobin/STAR/issues/376>



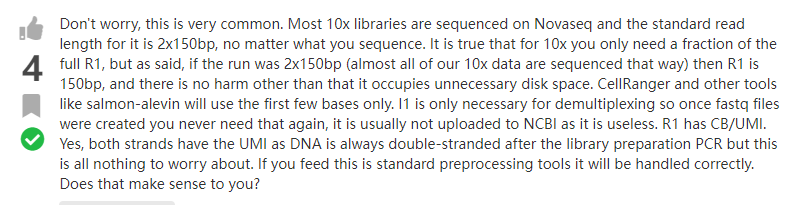
有人回复可能是用snakemake提交的时候预先设定了为程序提供的内存不够，因此修改cluster文件中相应rule的内存 —— 解决。【注意：cluster.yaml 中的条目必须和rule名字相同，否则默认为1】

1. **：使用MARVEL pipeline 使用的原始数据做分析，摸索pipeline略过的分析部分，以对比结果的方式判断分析是否有问题**。
2. 从参考文献找到原始数据sra号：SRR9008752、SRR9008753、SRR9008754、SRR9008755
3. 下载fastq文件： fastq-dump --split-files –origfmt SRR9008752，注意fastq-dump工具强烈建议不要添加 –gzip 参数
4. SRA提取fastq，并用gzip进行压缩

网址 <https://bioinformaticsworkbook.org/dataAnalysis/RNA-Seq/Single_Cell_RNAseq/Chromium_Cell_Ranger.html#gsc.tab=0> 展示了如何使用网络下载的fastq文件进行分析的流程。示例文件read1长度为102bp，我们使用的fastq文件read1为150bp，示例流程并没有说明read1的长度是否会对比对有影响。【已解决，参考问题13】

1. Cellranger提交比对
2. STAR提取Gene和SJ——需要明确barcode白名单和UMI长度（查阅原始文献，建库版本为V2，UMI长度8）
3. **：从sra直接提取的fastq双端reads都是150bp，数据无需要进一步进行trim【已解决】**

以 “any deal for fastq from sra for 10x genomics data of scRNA-seq read1 has 150bp”为关键词在网络上检索，找到网页 <https://www.biostars.org/p/9529864/#9557246> 回答了这个问题。如果使用cellranger或者salmon-alevin等10X提供的标准工具来处理，则无需担心任何问题，如果自行使用STAR比对，则需要指定barcode的位置和umi长度。



1. ：**比对时每个样本分别比对，后续多个样本是如何整合在一起的**？【**已解决**】
2. 10X数据可以通过cellranger aggr实现多个样本的数据合并，但cellranger只管基因，SJ要如何整合？如何保证如果cellranger和M20存在较大不同的时候保持一致？
3. M20数据尚未实现合并

10X和M20流程在比对的时候每个样本分别进行比对，以并行的方式增加流程运行效率，鉴于cellranger并不考虑SJ，其aggr也并不能整合SJ，因此，10X和M20都采用分别比对，然后使用seurat包merge函数合并样本。

samplelist <- c("SRR9008752", "SRR9008753", "SRR9008754", "SRR9008755")

min\_cells <- 3

min\_features <- 200

folder\_path <- "D:/temp/"

seurat\_list <- list()

for (sample in samplelist) {

  data\_path <- paste0(folder\_path, sample,"\_Solo.out/Gene/filtered/")

  seurat\_obj <- CreateSeuratObject(Read10X(data\_path), min.cells = min\_cells, min.features = min\_features)

  seurat\_obj$orig.ident <- sample

  seurat\_list[[sample]] <- seurat\_obj

}

**注意**：SJ矩阵的feature.tsv文件格式与gene不同，会报错，需要进一步处理一下，仅使用染色体:起始位置:结束位置表示SJ，其他信息按照格式做无效填充

1. ：**标准化和批次效应的关系**？**差异可变剪切的分析是否需要矫正批次效应？**【**已解决**】

网址 <https://blog.bioturing.com/2022/03/24/batch-effect-in-single-cell-rna-seq-frequently-asked-questions-and-answers/> 提供了批次效应和标准化的使用推荐。其中关于两者的定义：

* Normalization targets variance from sequencing (library preparation, high dropout event, amplification bias caused by gene length GC content, etc.) (Jia et al., 2017)
* Batch effect correction targets variance from experimental designs and handling (dfferent sequencing platforms, timing, reagents, laboratories, etc.) (Haghverdi et al., 2018)

Another noteworthy difference between normalization and batch-effect removal is in the input data. Specifically, while normalization works on raw count matrix (e.g. cells x genes), most methods to **remove batch effects use dimensionality-reduced data** (e.g., the first 50 principle components from Principal Component Analysis) to save computation time. This means that even though **the resulting batch-effect-removed result will be useful for visualization and graph-based clustering**, **other downstream analyses that require a full matrix of all genes (such as Differential Expression analysis) can not build up from the previous batch-effect removal process**. Exceptions for this drawback are methods like Mutual Nearest Neighbor (MNN) or scGen which will output a normalized gene expression matrix with the same dimensions as the raw count table (Tran et al., 2020).

差异表达分析不管批次效应。

1. ：**目前10X和M20的数据Feature文件中gene不唯一，明确处理方案**

在处理feature数据时，发现有多个基因存在一个gene\_name对应多个gene\_id的情况，这会引入数据丢失或者错误。

1. ：**从seurat对象转到MARVEL对象？【已解决】**

# 存储标准化的gene数据

seurat\_list <- list()

for (sample in samplelist) {

  data\_path <- paste0(folder\_path, sample,"\_Solo.out/Gene/filtered/")

  seurat\_obj <- CreateSeuratObject(Read10X(data\_path), min.cells = min\_cells, min.features = min\_features)

  seurat\_obj <- NormalizeData(object = seurat\_obj, normalization.method = "RC", scale.factor = 1e6 )

  seurat\_obj$orig.ident <- sample

  seurat\_list[[sample]] <- seurat\_obj

}

seurat\_merged\_normalization <-  Merge\_Seurat\_List(seurat\_list, add.cell.ids = samplelist, merge.data = TRUE)

# 提取矩阵数据

df.gene.norm <- as(as.sparse(seurat\_merged\_normalization@assays$RNA@counts), "dgTMatrix")

# 制作表型文件

df.gene.norm.pheno <- data.frame("cell.id" = row.names(seurat\_merged\_normalization@meta.data),

                                 "donor.id" = seurat\_merged\_normalization@meta.data$orig.ident)

# 制作基因文件

df.gene.norm.feature <- data.frame("gene\_short\_name" = unlist(seurat\_merged\_normalization@assays$RNA@data@Dimnames[1]))

# remove(seurat\_merged\_normalization)

remove(seurat\_obj)

1. ：**gtf文件格式问题：STAR（2.7.10b）使用的gtf文件和cellranger中较老版本的不同，在MARVEL的R脚本中读取现有的gtf文件报错？【已解决】**

**问题描述**：读取gtf文件时，使用从网络上下载的专门为10X数据准备的gtf文件就没有问题，但是使用我们自己为M20准备的gtf文件就报错，报错信息为：“检测到 1 个列名，然而数据共有 9 列（文件不合法）。添加了 8 个额外列名到结尾处”

**结论**：乌龙，下载示例文件不能用，咱们自己的文件可以用。

1. ：**MARVEL中注释不到基因信息？【已解决】**

GTF文件共9列，其中第九列包含基因注释的各项信息，如果GTF文件没有正确读取，后续分析就会报错。正确读取后为9列的数据框，需要重新给每列命名为V1~V9。

1. **：M20数据SJ矩阵过大，有否办法缩减一下？【已解决】**

STAR 说明书中有几个关于SJ的参数，

--outSJfilterReads 是否允许一个read有多个匹配，默认允许多重比对，我们设置为只允许单独匹配【参考问题36b)】

剩余几个参数都是和具体SJ判定有关，不能修改。

--outSJfilterOverhangMin、--outSJfilterCountUniqueMin、--outSJfilterCountTotalMin，--outSJfilterDistToOtherSJmin，--outSJfilterIntronMaxVsReadN

参数修改对于降低SJ矩阵的作用微乎其微，仅确保唯一匹配即可

【**最终解决方案**】

SJ矩阵过大，主要原因是所有的junction都是未经过滤的，feature特别多和cell都特别多，而算不动是因为在将Seurat对象转为MARVEL对象时，矩阵会占用过多的内存，从而算不动，解决办法为在将Seurat对象转为MARVEL对象前，就根据基因-细胞矩阵将SJ矩阵中高质量的细胞提取出来、再根据注释，仅报告成功注释到基因外显子的junction，就从两个维度大大降低了SJ矩阵的大小。

1. **：如何确定差异比较类型？样本间，cluster间，同一个cluster不同样本间？【已解决，可依据上游信息随意指定比较组】**

让比较组的判定在snakemake中完成，每次比较就重新投递一个R程序，而不是在R中做循环。

* 按照从简到难的原则：首先计算**样本间**的差异可变剪切

样本间的差异分析只需要在**元数据**中标记样本ID，在后续分析的时候以样本为单位提取信息即可。

【最终解决方案】

直接从上游的分类表中获取表头，所有在上游表格中存在的表头都可以用于后续分组，只需要将相应的信息放在config/diff\_group.csv文件中即可，下面为示例文件：

treatment,treatment\_name,control,control\_name,type

1,1,2,2,clusters

sham,sham,SCI\_1W,SCI\_1W,sampleid

B cells,B cells,Stem cells,Stem cells,celltype

NK cells,NK cells,Stem cells,Stem cells,raw\_main\_celltype

1. **：得到junction表达量、gene表达量的总表【已完成】**

Marvel\_analysis后就输出表达量总表了，以及pvalue、qvalue、在百分之多少的细胞中表达，总体平均表达量等信息都有。

1. **：绘制火山图【已完成】**
   1. 默认火山图没有legend，需要添加legend – 已完成
   2. SJ的筛选有log2fc和delta两种，但是默认图使用delta的时候是错误的，重新绘制，并实现在config文件中随意设置和传参 – 已完成
   3. 需要为特定junction添加注释
2. **：最低表达量参数设置【已完成，使用较低标准参数，给出尽可能多的结果给老师】**

整个marvel流程中有多个地方对样本平均最低表达量有筛选，其中对表达量的筛选在两个水平上实现：

1. 细胞水平：要求基因至少在百分之多少的样本中有表达。默认流程中会先绘制基因在百分之多少细胞中表达的密度图，然后找到密度图的最高点对应的基因表达量，只保留在最多细胞中表达的基因。
2. 在所有细胞中的平均表达水平。Marvel会统计每个基因在所有细胞中的平均表达量，这个值会直接从外部传参到流程中。

以marvel示例数据（Figure 1）为例：（A）显示junction表达量log2fc差异最大的junction主要集中在低表达量的基因上，高表达量的基因也存在部分差异较大的基因，若我们选择的最低表达量太低，就很难筛选到表达量高，同时还显著差异的junction。（B）显示Junction表达量delta差异值最大的junction全部集中在低表达量的基因上，且出现极值，这些基因很可能是10X数据覆盖度不高造成的随机现象，需要慎重对待。当把最低表达量调整为1时（即平均每个细胞中表达2个），以delta为纵坐标的图形（D）上就有很多候选基因出现了。比较（C）和（D）可以发现，对于高表达量的基因，delta参数具有更高的检测效力，比较（A）和（B）可以发现，对于低表达量的基因，log2fc具有更高的检测效力。

**建议**：在老师对候选的目的基因没有表达量的背景知识的时候，最好两种图都给老师绘制出来。

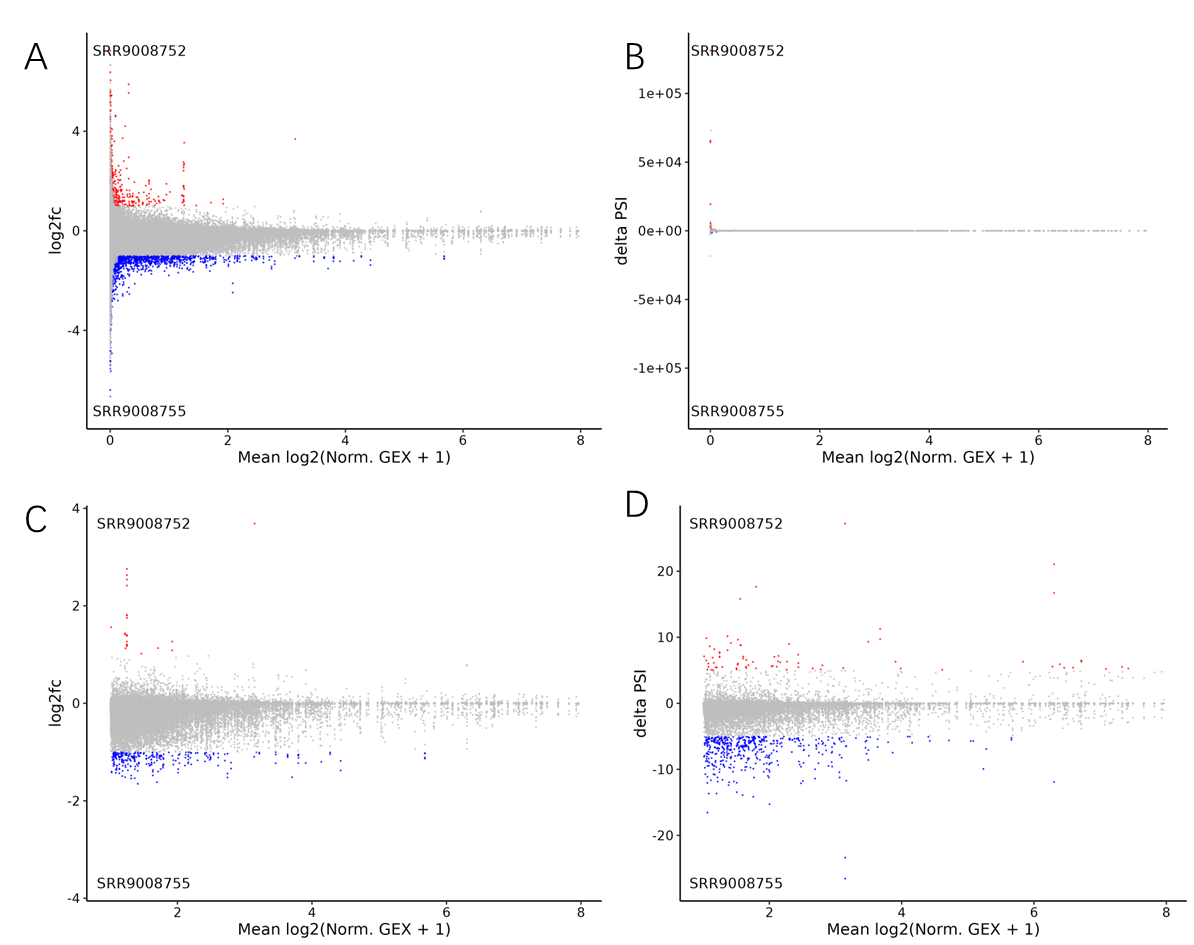


Figure 1．火山图。（A）以junction差异表达log2fc值为纵坐标，基因在所有细胞中的平均表达量为横坐标绘制火山图，最低表达量设置为-100，即保留所有基因；（B）以junction差异表达delta值为纵坐标，基因在所有细胞中的平均表达量为横坐标绘制火山图，最低表达量设置为-100，即保留所有基因；（C）以junction差异表达log2fc值为纵坐标，基因在所有细胞中的平均表达量为横坐标绘制火山图，最低表达量设置为1，即保留所有基因；（D）以junction差异表达delta值为纵坐标，基因在所有细胞中的平均表达量为横坐标绘制火山图，最低表达量设置为1，即保留所有基因。

1. **：差异junction的评估方式有两种：log2fc和delta，用哪个比较好？【已解决，delta更好】**

参考问题26，对于高表达量的基因，delta参数具有更高的检测效力，对于低表达量的基因，log2fc具有更高的检测效力，具体使用哪一种，首先看老师的先验知识，最好全都绘制，为老师提供更多的选择。

1. **：差异基因的筛选使用pvalue和qvalue? 【已解决】**

常规单细胞分析差异基因使用pvalue，因此在可变剪切基因表达量差异统计的时候也是用pvalue作为默认的筛选标准，同时提供传参，方便进行选择。【注意：只是基因的差异表达有qvalue，junction表达量是否存在差异是通过permutation得到的，没有qvalue】

【为什么会存在qvalue？在什么时候推荐使用qvalue?】

基因差异表达使用秩和检验，每个基因单独比较，多重比较的时候会引入更多的错误，需要使用q-value进一步矫正，但是junction的差异分析使用的是permutation，此时不需要使用q-value。

1. **：可变剪切的类型统计【已完成】**

基因的差异表达分为两种，分别是差异上调和差异下调，但是可变剪切的差异要比基因表达差异复杂的多。

1. 基因只检测到一个差异表达的junction，差异方向可能和基因差异方向相同，也可能相反，也可能基因没有差异表达
2. 检测到两个及以上差异表达的junction，不同junction的差异方向可能相同，也可能相反，和基因表达差异的方向也可能相同，也可能相反，也可能基因表达量并没有差异。

Marvel将按照junction和基因表达量差异方向是否相同将基因分为4中类型：

* Coordinated：junction的差异方向和基因方向相同
* Opposing：junction的差异方向和基因方向相反
* Iso-Switch：junction表达量有差异，但是基因表达量没有差异
* Complex：基因表达量有差异，且有多个junction，且junction差异方向不同

考虑到部分基因并没有差异表达，但是至少两个junction表达差异方向相反的基因也很有可能是老师想要的可变剪切候选基因，因此我们将 Complex 进行扩展：

* New Complex：不考虑基因表达是否存在差异，只要一个基因具有至少两个junction，且存在差异方向相反的junction，就认为还基因是Complex 基因。

**可变剪切首先需要有至少两个junction，因此只有一个junction的基因只保存在表中，并不进行绘图展示。其中 Complex 基因是最重要的候选可变剪切基因**。

1. **：可变剪切基因的筛选标准？【已完成】**

参考问题27，Complex基因是最重要的可变剪切候选基因，在绘图时全部展示，其他类型的候选基因，只挑选一个基因含有多个junction的，且junction表达差异最大的前几个基因进行绘图。

目前候选基因筛选时除了p-value、delta等标准参数外，按照基因差异表达方向和junction差异表达方向将候选基因分为四类：

* Coordinated: 基因所有差异表达junction的差异方向和基因方向相同；
* Opposing: 基因所有差异表达junction的差异方向和基因方向相反；
* Iso-Switch: 基因差异表达junction差异方向相同，但是基因没有差异表达；
* Complex: 基因同时包含2个及以上差异表达junction, 且存在差异方向相反的情况；

其中 Complex 类型的基因最有可能检测到可变剪切。

1. **： junction面板图型绘制逻辑修改。【已解决】**

marvel默认流程在绘制junction的面板图时默认将所有junction都绘制出来，但是单细胞数据很难将所有junction都检测到，且并不是所有junction都差异表达，因此我们修改逻辑，默认情况下只绘制候选的差异表达的junction，但是仍然保留参数接口，让流程可以绘制所有junction。

需要给每个junction加上label。并把对应表格输出出来。

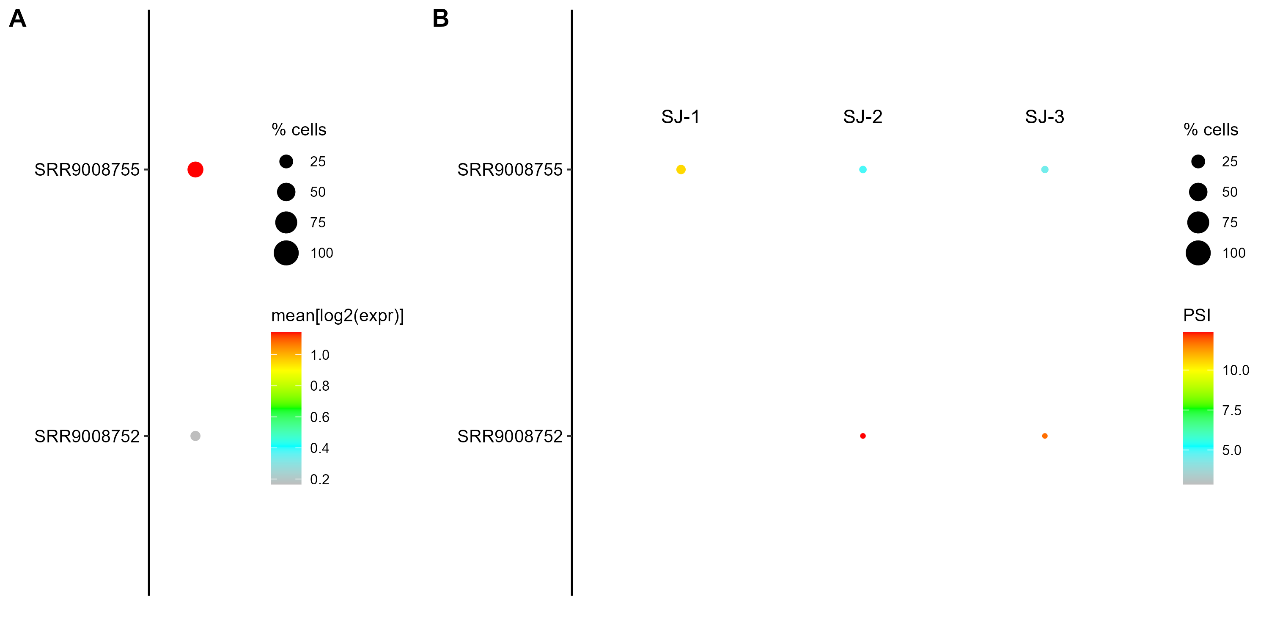
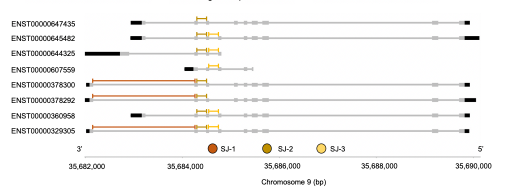


Figure 2．基因和junction表达量面板图。（A）基因表达量；（B）差异表达的junction表达量。

1. **： 基因结构图绘制：一方面pipeline没有提供有多个junction时的绘图代码，另一方面绘制单个junction的源码已不能使用**

Marvel流程提供了基因结构的示意图，如下图所示：



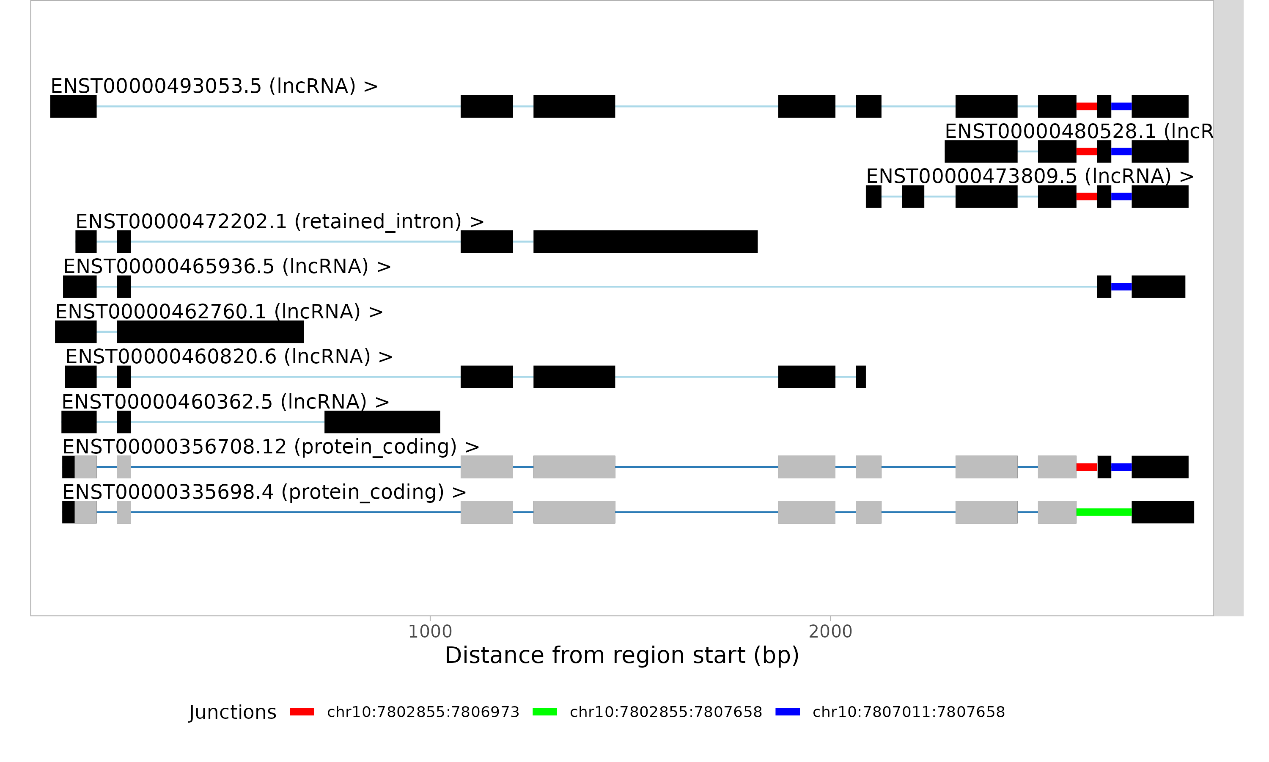
鉴于问题中描述的内容，我们需要自行更新绘图代码。按照源码的基本思路：从gtf文件中将所有的转录本信息都提取出来，所有包含该junction的转录本都绘制一段彩色的junction（线段）进行展示。

**思路1**：marvel源码中将所有的junction制作成假转录本，并将junction作为一个独立的转录本进行绘制。我们再此思路的基础上进一步将多条junction信息同时添加到进来，并同时进行绘制，再从ggplot对象中将junction提取出来，制作新图层，为junction添加颜色

**结果1**：该方法遇到错误，当junction多于1个时，图可以正常绘制，但是图片上的信息是错误的，一行一行查看代码，发现错误是从调用绘图工具wiggleplotr::plotTranscripts开始的。查看wiggleplotr源码，发现junction信息被错误的合并了。

**思路2**：不再次用marvel原始的处理思路，不再将junction制作为假转录本，单独绘制所有基因的结构，再在当前结构的基础上，从比对信息和注释信息中找到哪些转录本含有junction，再增加图层，给junction标记线段。如果设置rescale\_intron为true，则从按照位置找junction对应intron的位于哪两个exon之间，从第一图层的数据中找到转换后的junction位置。

**结果2**：思路2成功解决了该问题，新的结构图片展示如下：



1. **：表达量面板图和基因结构图中的junction需要一一对应【已解决】**

并非都用颜色区分junction，无需对应

1. **：10X 数据 SJ 缩减，降低资源消耗【已解决】**

SJ矩阵为未过滤的矩阵，所有检测到的细胞都包含在内，在将样本合并后，我们可以根据基因矩阵中的细胞信息从SJ矩阵中提取细胞，一次性降低SJ矩阵的大小。【现在的处理方式和问题33保持一致】

1. **：M20数据量过大， SJ矩阵处理困难，要怎么缩减矩阵？【已解决】**

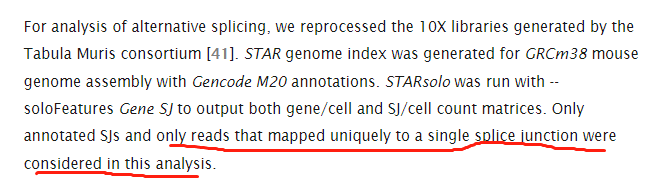
M20的SJ矩阵要比10X数据的SJ矩阵大的多，问题32中的解决办法无法解决本问题，还需要进一步降低矩阵大小。

思路：read10X函数读取数据后存储格式为dgCMatrix，可以直接当做矩阵处理。SJ矩阵中包含所有的junction，但是在后续分析时我们仅使用完整匹配外显子的junction；除提取junction外，再根据基因矩阵的barcode信息，从SJ矩阵中提取细胞。此时SJ矩阵就会降低很多，再将矩阵转换为seurat对象。

1. **：10X数据和M20数据在检测到可变剪切上有显著差异么？【搁置】**

由于目前M20数据质量问题，该问题暂时搁置

1. **：能否用junction进一步对cluster进行降维，从而鉴定细胞亚型。**
2. **：sashimi plot绘图，如何整合bam文件。**
3. 工具是python2，需要改成python3
4. 使用bam文件为输入文件，但是目前bam文件是以样本为单位存储的，需要拆分为以细胞为单位进行存储的形式
5. Bam文件中包含所有的 reads信息，需要把比对过程中过滤掉的低质量reads去掉
6. **：其他需要注意的问题**
   1. GTF文件中 gene\_name 不能有下划线 【有项目正在进行，暂时没有修改】
   2. 虽然MARVEL pipeline 中并没有指定参数 --outSJfilterReads Unique， 但是文献中提到 only reads that mapped uniquely to a single splice junction were considered，因此添加该参数。



* 1. 可变剪切只要有junction就可以了，junction在两个样本间差异表达
  2. 使用yaml直接传参，不再使用0.common.smk进行传参，节省大量时间精力，同时也使代码更加整洁

results$log2fc <- log2( (results$psi.g2 + 1) / (results$psi.g1 + 1) )

results$delta <- results$psi.g2 - results$psi.g1